View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital Repository @ Iowa State University

IOWA STATE UNIVERSITY
Digital Repository

Graduate Theses and Dissertations

Iowa State University Capstones, Theses and Dissertations

2006

Modeling crash frequency and severity using historical traffic and weather data: truck involved crashes on I-80 in Iowa

Micah Makaiwi
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

 Part of the Civil Engineering Commons, and the Transportation Commons

Recommended Citation

Makaiwi, Micah, "Modeling crash frequency and severity using historical traffic and weather data: truck involved crashes on I-80 in Iowa" (2006). *Graduate Theses and Dissertations*. 15056.
https://lib.dr.iastate.edu/etd/15056

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Modeling crash frequency and severity using historical traffic and weather data: Truck involved crashes on I-80 in Iowa

by

Micah Makaiwi

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Civil Engineering (Transportation Engineering)

Program of Study Committee:

Jing Dong, Major Professor

David Cantor

Peter Savolainen

Iowa State University

Ames, Iowa

2016

Copyright © Micah Makaiwi, 2016. All rights reserved.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

As congestion grows along roadways in the country, it is important to see how this will affect crashes on America's highways. I-80 in Iowa is a major trucking corridor for transferring goods between the east and west coasts and carries an increasing volume of freight trucks on the road. The recent ability to record detailed speed and volume data over Iowa's road system presents a new opportunity to examine whether congestion and slowdown affect the occurrence and severity of crashes along I-80. This study examines the use of INRIX speed data, Wavetronix radar data and Road Weather Information Systems [RWIS] data on I-80 in Iowa to model freight truck crashes. A random-parameter Poisson regression model is used to examine how speed, weather and roadway characteristics affect the frequency of crashes along different segments. An ordered probit model examines how these factors affect the severity of injuries in truck crashes. In general, lower speeds and congestion were associated with more frequent crashes (taking into account the vehicle-miles travelled) of lower severity. High speed, low congestion periods are more often associated with fewer, but more severe, crashes.

CHAPTER 1. INTRODUCTION

Interstate 80 is a vital freight transportation corridor crossing the middle of the state of Iowa and passing through some of the largest metropolitan areas in the state. It is one of the major corridors connecting the east and west coasts and much of this traffic passes through Iowa. The 301 mile long corridor connects some of the largest metropolitan areas in the state, such as Council Bluffs, Des Moines and the Quad Cities, and provides many connections to other modes of freight transport in the state.

However, this important corridor will experience increasing traffic congestion in the future as traffic volumes increase faster than capacity. According to the 3rd edition of the Federal Highway Administration's Freight Analysis Framework, 3.8 miles of I-80 in Iowa had peak-hour speeds half of free-flow speed or worse in 2007; in 2040, that figure is projected to increase to 115.5 miles (Federal Highway Administration, 2011). It is important to know how this increasing congestion will affect traffic safety along I-80.

Recent advances in traffic counting and data storage have made it easier to collect detailed data on traffic volumes and speeds. The Iowa Department of Transportation [Iowa DOT] has access to two data sources that are of particular interest: INRIX speed data and Wavetronix radar data. INRIX derives traffic speed and travel times for much of the United States using GPS controllers in commercial vehicles, taxis and personal cell phones. The INRIX data has wide continuous coverage over almost all highways and arterial in the United States. In addition, the Iowa DOT has placed Wavetronix radar sensors at strategic locations throughout Iowa. These stations use radar to measure traffic volumes and speeds for the vehicles at the sensors for each lane. Unfortunately, after examining the Wavetronix dataset, it was determined that the stations do not have enough

coverage to get a large enough sample of crashes for analysis. It was, however, used to validate and calibrate the measurements obtained from other sources as shown in chapter 3.

Truck crashes are a particular concern. Trucks are larger and less maneuverable than passenger cars. Other concerns include driver fatigue; truck drivers work long hours driving across countries and often face pressure from clients and employers to work long hours and drive excessive speeds.

The objective of this thesis is to look at freight crashes from the Iowa DOT crash database and to geospatially and temporally relate these crashes to the INRIX and other datasets including weather and automatic traffic recorders [ATR]. Since multiple data sources were used, a linear referencing system [LRS] was developed using geographic information systems [GIS]. Each dataset was put into the LRS to calculate where along I-80 each record is located. This converted the two-dimensional geospatial data into a single number that could be compared to other data sources simply and accurately.

To examine how traffic and environmental characteristics affect crash frequency and severity, two models were developed. The first is a crash frequency model, which used a random-parameter Poisson regression to calculate the likelihood of a crash on a particular segment in a particular month. I-80 was segmented by the Traffic Message Channel [TMC] segments present in the INRIX dataset. Data from the Iowa DOT's Geographic Information Management System [GIMS] roadway network were combined with the DOT's crash database, the road weather information system [RWIS] and automatic traffic recorder [ATR] data. Since this is a random-parameter model, some of the coefficients from the regression were allowed to vary randomly by TMC segment. The results show that higher traffic volumes (measured by vehicle-miles traveled [VMT]), the percent of time in a month that the roadway was icy, and the percent of trucks on the roadway lead to higher crash frequency. Wider shoulders and the months December and January corresponded to lower number of crashes. In general, slower speeds—which indicate traffic congestion—increased the frequency of crashes. Two speed-related variables from INRIX were used: the percent of time that traffic was going at the speed limit or faster and the percent of time that traffic was going slower than 10 mph below the speed limit.

The second is a crash severity model using an ordered probit model, which helps determine what factors affect the likelihood of a crash having injuries or fatalities. Each crash was a record in the model and the severities were grouped into three groups by the most injured person in the crash: fatal and major injury crashes; minor and possible injury crashes; and property-damage only crashes. This model found that higher speeds in the 30 minutes prior to the crash led to more severe crashes as well as crashes involving multiple trucks, crashes involving a non-truck vehicle, crashes where a truck rear-ended a vehicle, crashes caused by swerving, run-off-road crashes and crashes where drugs or alcohol was involved. Crashes where the road was chemically wet or covered in snow/ice, crashes on the weekend, and sideswipe crashes were associated with less severe injuries.

Chapter 2 details literature about crash frequency and crash severity models as well as some common issues with them. This is followed by an examination of current research into truck-related crashes. This is followed by chapter 3, which details the different data sources used in the analysis: Wavetronix, INRIX, the Iowa DOT crash database, GIMS, the ATR recorders, and RWIS. Chapter 4 details how these different data sources were associated to each other and condensed down into two different data sets for the crash frequency and crash severity models, whose results are presented in chapter 5. There, all the results and the findings from the model are presented. Finally, everything is summarized in chapter 6.

CHAPTER 2. LITERATURE REVIEW

While much research has been performed on traffic crashes in general, there is much less insight into many of the causes and contributing factors in crashes involving trucks. Trucks have unique operating characteristics different from other road users. Trucks are larger; this makes trucks more difficult to maneuver and make impacts more deadly.

Truck drivers often drive long distances and have financial incentive to drive long hours. The Federal Motor Carrier Safety Association [FMCSA] regulates the number of hours truck drivers can drive in a particular time period. As of March 2015, FMCSA rules state that drivers can drive at most eleven hours at a time with at least a ten hour break and must have taken a thirty minute break within the last eight hours. (Federal Motor Carrier Safety Administration, 2011)

Kraft et al. (2009, p. 16) provide a detailed analysis of the aspects of trucks that impact the way they use roads. Driving a truck is much more complex than driving a passenger vehicle due to the increased size and weight. In general, truck drivers require a special commercial drivers license to operate their vehicles. They also accelerate slower which means that they need larger gaps on freeways to merge and cause more disruptions to the traffic stream during congestion.

Mason and Smith (1988) note that there are many traffic control devices and regulations that only affect truck drivers. In addition, trucks have the potential to jackknife and are much more susceptible to rollovers and high wind. They generally travel slower than other vehicles and are frequently passed which is compounded by larger blind spots behind the vehicle.

2.1 Crash frequency models

A common traffic crash model is a count or frequency model. A count model takes a particular segment of roadway and determines what factors are associated with a specific number of crashes occurring during a time period.

A traditional linear regression is inappropriate for this situation. Count data are not generally normally distributed and they often have a skewed distribution. In addition, counts cannot be negative. To combat these issues, the Poisson distribution is used which much more properly describes count data. Traditional linear models predict the expected value of Y given x, $\mathrm{E}[Y|x]$, given the sum of multiple coefficients β_i multiplied by the matrix of predictor variables, x. A Poisson regression model is a generalized version of the standard linear regression model. It transforms the typical linear model, $\mathrm{E}[Y|x] = \sum \beta_i x_i$ using a logarithm so that $log(\mathrm{E}[Y|x]) = \sum \beta_i x_i$.

However, the Poisson distribution has one major assumption that is often violated by crash data: the mean of the dependent variable ($\mathrm{E}[\mu]$) is approximately equal to its variance $\mathrm{VAR}[\mu]$. When this is not satisfied, the data is considered overdispersed ($\mathrm{E}[\mu] < \mathrm{VAR}[\mu]$) or underdispersed ($\mathrm{E}[\mu] > \mathrm{VAR}[\mu]$). The negative binomial regression model is an extension of the Poisson regression model that accounts for over-dispersion or under-dispersion with a dispersion factor α. The α value is a coefficient allows the model to account for when the variance does not equal the mean.

Lord and Mannering (2010) lists many other models for crash frequency. These include the gamma model, bi-variate/multivariate models and slight modifications to the Poisson model—such as a Poisson-log-normal model or a Conway-Maxwell-Poisson model. Another alternative analysis is duration models, which model the duration between crashes. While this model does not handle time-varying data well, it is a good way to handle the rarity of crashes and low mean number of crashes.

According to Lord and Mannering (2010), the most common crash frequency model is to use the overall number of crashes in one crash frequency model and deal with injury severities after

determining the total number of crashes. Some research uses separate crash frequency models for different injury severities, but there is often a lot of correlation between the individual models and requires a complex model structure to account for this correlation.

Another formulation of the count model is the zero-inflated count model. As discussed by Lord et al. (2007), the zero-inflated Poisson [ZIP] and zero-inflated negative binomial [ZINB] models have fallen out of favor from many experts in transportation safety. Zero-inflated count models assume that there is some portion of the population that will inherently have a count of zero. They argues that the transportation system never is in a "true-zero" or "virtually safe" state.

2.2 Injury severity models

Much of the analysis of crashes has been finding what factors affect injury severity in crashes. The ordered probit and logistic regression models are the most frequently used models for injury severity due to the inherent ordered nature of crash injuries. The typical scale used by most agencies in the United States in the KABCO scale. This scale contains five injury severities, each with an associated letter. The Iowa Department of Transportation (2014) uses a numeric code from 1–5, with one being the most severe. These are defined below with the Iowa DOT code and the corresponding KABCO code:

- **K/1–Killed/fatal injury:** "used when a fatal injury is any injury that results in death within 30 days after the motor vehicle crash in which the injury occurred. If the person did not die at this scene, but died within 30 days…the injury classification should be changed…"
- **A/2–Major/incapacitating injury:** "used when any injury, other than a fatal injury, that prevents the injured person from walking, driving, or normally continuing the activities the person was capable of before the injury occurred. This includes severe lacerations…broken or distorted limbs…significant burns…unconsciousness at or when taken from the crash scene.…This does not include momentary unconsciousness."
- **B/3–Minor/non-incapacitating injury:** "used when a minor injury is any injury that is

evident at the scene of the crash, other than fatal or serious injuries. Examples include lump on the head, abrasions, bruises, minor lacerations (cuts on the skin surface with minimal bleeding and no exposure of deeper tissue/muscle. This does not include limping."

- **C/4–Possible (complaint of pain/injury):** "used when a possible injury is any injury reported or claimed that is not a fatal, suspected serious, or suspected minor injury. Examples include momentary loss of consciousness, claim of injury, limping, or complaint of pain or nausea. Possible injuries are those that are reported by the person or are indicated by his/her behavior, but no wounds or injuries are readily evident."
- **O/5–Property damage only [PDO]/uninjured:** "used when there is no apparent injury and there is no reason to believe the person received any bodily harm from the motor vehicle crash. There is no physical evidence of injury and the person does not report any change in normal function."

In addition, the Iowa DOT recognizes two other injury severity levels for individuals involved in a crash. These injuries are grouped with a category above for the purposes of determining the overall crash severity.

- **7–Fatal, not crash related:** "used when the vehicle fatalities that are involved in a motor vehicle crash have died from natural causes such as a stroke, heart attack, or from a homicide or suicide." *Grouped with A/Major Injury crashes*
- **9–Unknown:** "used when the person has left the scene and is unknown" *Grouped with C/possible injury crashes*

However, the boundary between some severities (particularly between the three injury severity levels) can vary between agencies and even between different officers. The severity of the crash is the most severe injury sustained in the crash. In Iowa, a crash injury is considered fatal only if a passenger in the crash died of injuries within 30 days of the crash. In cases where the death was not caused directly by the crash, such as a heart attack, the crash is counted as major injury. In

addition, Iowa also has another injury severity level, "unknown." Individual unknown injuries are reported and crashes with only unknown injuries are grouped with possible injury crashes.

As seen above, there's a strong ordered relationship between the different injury severities. There are many ways to analyze discrete-choice ordered data. Savolainen et al. (2011) describe how these models are applied in a transportation context and some of their methodological advantages and shortcomings. While there is an inherent ordering to the values (e.g. fatalities are more severe than minor injuries), the values are categorical in nature and not numerical. Unlike continuous data—which is also ordered—each discrete category does not have a numerical value that mathematical operators cannot quantify; the difference between a fatal and a major injury crash is not necessarily numerically quantifiable. In addition, a model for continuous variables such as a standard linear regression can predict values that are outside the range of allowable answers. There are methods that assign a monetary value to specific injuries but this is not always the most applicable to statistical analysis.

The most common models for ordered discrete data are the ordered logit and probit models. These use a generalized linear model that predicts the value of an exact, but unobserved continuous value, z, where z follows a typical linear regression: $z = \boldsymbol{\beta X} + \varepsilon$ where $\boldsymbol{\beta}$ is a vector of estimable coefficients, $\boldsymbol{X}$ is a matrix of predictor values and ε is a normal disturbance term. To determine which discrete variable the model will predict, a variety of thresholds μ_i are developed to predict the final ordinal predicted value y as shown in equation (2.1)

$$y = \begin{cases} 1 & \text{if} \quad z \leq \mu_0 \\ 2 & \text{if } \mu_0 < z \leq \mu_1 \\ \dots & \\ i & \text{if } \mu_{i-1} \leq z \end{cases} \tag{2.1}$$

There are many ways to formulate ordered severity models. One common method is to group different severity levels together (often grouping all injury crashes together or grouping fatal and incapacitating injury crashes). This allows for sufficient observations in each group since the number

of highly severe crashes often is much lower than property-damage only. In addition, grouping so there are only two groups (e.g. injury and non-injury) allow for analysis with a binomial model such as a logistic regression.

Another approach sometimes seen is a multinomial logit model. This model does not account for the inherent order of the severity levels but allows for extra flexibility. For instance, an ordered logit/probit model forces regressors to have the same coefficient across all severities. This may not always be the case. A multinomial logit model allows a regressor to have different effects across different severity levels; however, these models do not take into account the inherent ordering of crash severity.

2.3 Goodness of fit measurements

Both types of models described above, the Poisson/NBL models and the ordered probit and logit models are a form of generalized linear models, which can be estimated using maximum likelihood estimation [MLE]. The MLE process produces a likelihood value; this is often logged to produce a log likelihood value. Better fitting models produce larger log likelihood values. Most goodness of fit measurements used in this study relate to these log likelihood values.

In general, the log likelihood values are extremely dependent on the dataset used. In order to determine a baseline value to compare the model to, a restricted log likelihood is calculated from the corresponding model with only a constant term included. With traditional linear regression models, the R^2 value is used to determine what proportion of the variance in the dependent variable the model describes. While the R^2 value is not applicable to other regression models, various pseudo-R^2 metrics have been devised. The one used in this study is the adjusted McFadden's pseudo-R^2. (Washington et al., 2010) This measure penalized the test statistic based on the number of variables included in the model because adding new variables will always improve statistical fit, regardless of if they are significant. The formula for the adjusted McFadden's pseudo-R^2 is

$$R^2_{adj} = 1 - \frac{LL_{full} - K}{LL_{restricted}}$$

where LL_{full} is the log likelihood of the model with the full set of regressors, $LL_{restricted}$ is the restricted log likelihood (the log-likelihood of a comparison model, usually constant only) and K is the number of regressors.

Another test to determine whether a model is significantly better than another is the likelihood ratio test. Washington et al. (2010) This test statistic is based on the log of the ratio of the log likelihoods of the two models being compared. It follows a χ^2 distribution, where a large test statistic and a low p-value suggests that the alternative model is more significant than the base model with the degrees of freedom being the difference in the number of parameters between the two models. The formula for the likelihood ratio test statistic for comparing a base model to an alternative is

$$D = 2\left(LL_{full} - LL_{restricted}\right)$$

2.4 Common issues with crash models

Due to the nature of crash data, it is difficult to produce an experimental setup with a controlled environment, so most studies—including this one—instead rely on police reports from crashes. This can lead to many violations of statistical assumptions. One major issue that affects both crash frequency and crash severity models is the underreporting of crashes. Blincoe et al. (2002) found that 25% of minor injury and up to 50% of PDO crashes were unreported. Ye and Lord (2011) looked into how underreporting of crashes affects common crash models. They showed that underreporting of crashes causes models to have an increased root mean square error [RMSE]. They recommended that for multinomial logit models and mixed logit models have fatal crashes as the fixed case and that ordered probit models should have crashes ranked in descending order from fatal to PDO.

The Poisson regression loses much of its power and often can have biased estimates when the mean of the dependent variable is low. The low-mean problem for Poisson regression models has been discussed by many authors recently. Wood (2002) formulates a method to determine whether it is likely that the model does not fit the data well by grouping individual records, taking the mean of them and then developing a G^2 test statistic using the following formula:

$$G^2 = 2\sum_{i=1}^{n} r_i \left[log\left(\frac{\bar{y}_i}{\hat{\mu}_i}\right)^{\bar{y}_i} - \bar{y}_i + \hat{\mu}_i \right]$$

where r_i is the number of records that were averaged to produce that group, $\bar{y}_i$ is the average predicted value in the group and $\hat{\mu}_i$ is the calculated predicted value based on the mean of the x values. This test statistic is compared to the χ^2 distribution. If the G^2 is greater than the critical χ^2 for a given level of significance, then it is likely that the model does not fit the data well. Unfortunately, as the sample size grows, this test statistic grows too so it is more difficult to diagnose problems with large sample sizes.

Crash data often have correlation between observations. For instance, crashes can be correlated spatially (occurring in the same area of a roadway) and temporally (crash patterns tend to vary across time). One way to account for this in a model is with panel data: data collected across the same observational unit (such as a roadway section) (Lord and Mannering, 2010). One way to improve any linear or generalized linear model's fit is to use a random-parameter or fixed-parameter model (Washington et al., 2010). These models remove the assumption that coefficients are constant values and instead let the coefficients vary across observations using a defined statistical distribution. This can account for unobserved heterogeneity in the model. The coefficients can vary for every observation or can vary by across a grouping of observations in panel data. A similar formulation is a fixed-effects model, which estimates a different constant coefficient for each group in panel data.

2.5 Crash-related studies

There have been many studies of the factors that affect crash frequency and severity. This section highlights many of the studies that analyzed similar variables to the one in this study. While there are many studies focusing on crashes, less work has been done with truck-related crashes. Fortunately, many of the factors that contribute to crashes should be similar between truck-involved crashes when compared to the overall population of crashes.

One way to account for different traffic volumes is to segment the models based on the time of day. This was done by Pahukula et al. (2015) by examining the crash severity of truck crashes on urban freeways. They ran five separate random-parameter muitinomial logit models split by the time of day. The injury severities were grouped by severe/fatal injuries, injury crashes and non-injury crashes. Only four variables were included in every model: restraint use (seat belts/helmets), male drivers, drivers younger than 35 and sideswipe collisions. However, each of these variables had different effects depending on the time of day; for instance, depending on the time of day, restraint use either led to an increase or decrease in the likelihood of a major injury occurring in a crash. Some were consistent, however; drivers under 25 were less likely to be involved in PDO crashes at all times throughout the day.

Zhu and Srinivasan (2011) provided one of the most detailed examinations of the factors that affect truck crash injury severity. Two different data sources were compared; both data sources were from the Federal Motor Carrier Safety Administration's [FMCSA] Large Truck Crash Causation Study [LTCCS]. The first data set was based on the severity of the crash as determined by the police officer accident reports. The second was based on the injury severity determined by the LTCCS researchers. This data source was augmented by finding more detailed information about many of the "human factors" in the crash that are not typically gathered by police reports. The results between the two models were often contradictory. Major findings are that wet roads and weekdays tended to have a lower injury severity while sideswipes, head-on collisions and higher speed-limit roadways had higher injury severity. Smaller sized trucks had lower severity and younger drivers

were more likely to be involved in a severe crash. However, many of the variables in this study had many records with unreported driver behavior variables which may have been endogenous (when drivers have incapacitating injuries the police officer often can't get as much information from the scene). Chen and Chen (2011) found that weekday crashes are more likely to be non-incapacitating injury crashes compared to other injury severities using a mixed logit model.

Martin's study (2002) took a look at the relationship between traffic volume and crash rates. This study used a negative binomial regression against the crashes per vehicle mile. They found that the incidence rate of property-damage only crashes and injury crashes was the highest when traffic was light, but the absolute number of crashes during high traffic periods was higher due to increased exposure. Heavier traffic had a lower rate of fatal crashes and weekdays had a higher crash rate than weekends.

Stein and Jones (1988) applied a methodology that differs from most crash studies called the case-control method. Whenever a crash occurred in the study area, three trucks would be selected at random at the same time and place a week later for a survey. These surveyed trucks were compared to the trucks involved in the crash to determine the relative frequency of different truck configurations. They found that large and double-trailer trucks were significantly overinvolved in crashes as well as crashes by young drivers and empty trucks.

The collisions in truck crashes often have different characteristics than between passenger cars. Duncan et al. (1998) found that in truck rear end crashes, crashes are more severe when the passenger car is rear ended instead of the truck. The interaction of cars being struck in the rear and speed differential was also very statistically significant. Golob et al. (1987) also found that rear-end crashes were more dangerous than other types but they did not find significant differences between trucks rear-ending passenger cars and vice versa.

Dong et al. (2016) used a zero-inflated negative binomial to determine what roadway characteristics affect truck crash severity. They found the largest effects came from the percent trucks; the higher the percent, the more crashes are likely. In addition, higher AADT, longer segment length, and a higher speed limit were associated with higher crash frequency.

Islam and Hernandez (2013) used a random-parameter ordered probit model to find factors that impact crash severity for crashes involving large trucks on U.S. interstates. This study defined a large truck as any truck with a gross vehicle weight rating [GVWR] greater than 10,000 pounds. The random parameter model used helped to explain much of the unobserved heterogeneity when compared to a fixed-parameter model. The study showed that curved highway sections, summer months, run-off-road crashes, speed-related crashes and crashes involving truck drivers from Texas were more likely to have injuries or severe injuries. Weekends, multi-vehicle crashes, trucks getting rear-ended, sideswipe crashes, rollover crashes, use of restraints and male occupants were associated with lower injury severity.

A fixed-parameter model was run to compare with the random-parameter model. The fixed-parameter model was found to be much less significant than the random-parameter model. In addition, the random-parameter model addresses some of the issues with an ordered probit model as described in section 2.4. For instance, a random-parameter model allows for some variation of the effects between different severities in the model. The article uses airbag deployment as an example. Airbag deployment reduces the likelihood of a fatality in a crash, but can increase the change of a minor injury from the airbag itself. Random-parameters allow the effects to change in magnitude and sign between different observations.

Another analysis of truck-related crash injuries was performed by Islam et al. (2014). Instead of an ordered model, this study used four separate mixed logit regression models for four scenarios (single-vehicle urban, single-vehicle rural, multi-vehicle urban and multi-vehicle rural). For each logit model, the severities were collapsed into three groups, fatal and major injury crashes (K and A), minor injury crashes (B), and possible injury and property-damage only crashes (C and O).

The study found different effects for each of the four scenarios, but some variables had consistent effects. For instance, it was found that in all urban crashes, trucks with a gross weight greater than 26,000 pounds had increased probability of being in injury crashes. The study found that off-peak times led to an increase in crash severity in rural multi-vehicle and rural single-vehicle crashes and that the PM peak was associated with an increase in probability of possible/no injury crashes in

the rural single-vehicle model. Crashes where a vehicle struck a fixed object were associated with an increase in major injuries for both rural models, with a larger effect for single-vehicle crashes.

CHAPTER 3. DATA

This thesis integrates many data sources together. This chapter details all of the data sources that were explored for this analysis as well as any issues and weaknesses discovered in the data sources. All the data was compiled into a relational database and relevant geospatial data were included in this database for analysis in ArcGIS. Refer to chapter 4 for the methodology of associating all the data sets.

3.1 Wavetronix traffic radar detectors

The Iowa DOT in recent years has been placing radar detectors produced by Wavetronix along interstates and major highways in Iowa. The majority of detectors are in the state's major metropolitan areas and provide many benefits for the DOT including incident management and traffic planning. These stations use radar to count vehicles, classify them and register traffic speeds. Table 3.1 details all the variables in the Wavetronix dataset. Many records in the database are incomplete and only have some of the fields populated. For instance, more records have vehicle counts per-lane, per-vehicle class thatn per-vehicle class only. The sensors also determine an aggregate vehicles per hour [VPH] count. Historical data from these sensors are available starting in September 2012, with more sensors becoming available through the years. Sensors are located on major highways and interstates in Council Bluffs, Sioux City, Des Moines, Ames, Iowa City, Cedar Rapids, and Davenport.

Most sensors on roadway mainlines record traffic in both directions. Ramp sensors generally only report one-way flows. Figure 3.1 shows maps of all of the sensors deployed by the Iowa DOT and all the ones along the mainline of I-80. The availability of data varies by stations. Through the

Table 3.1: Fields in the Wavetronix dataset

Group	Fields	Availability[a]
Basic count data	Raw count, VPH, occupancy, speed	96.2%
Data quality	VPH quality, occupancy quality, speed quality	95.3%
Count by vehicle class	Counts of vehicle classes 1–4	38.1%
Count by lane	Count, VPH, occupancy, speed for lanes 1–8	84.1%
Count by lane and vehicle class	Counts of vehicle classes 1–4 for lanes 1–8	69.6%

a: Availability is the percent of the records in the dataset containing these fields

analysis period, new stations were added and each city received its first station at different times. Figure 3.2 shows when each station in the network was operational (moving aling I-80 from west in Council Bluffs, east to Davenport). For each station, months where it was operational are marked in figure 3.2

3.1.1 Wavetronix data accuracy

Since the Wavetronix data are very new, there are some potential issues that must be addressed. The major concern is that the availability of the data are not consistent across stations. Figure 3.2 shows the availability of records for each station, aggregated by month in the analysis period. Some stations have periods in the middle where they do not have any data and each city has inconsistent times when the stations were turned online. So depending on the timeframe, different crashes on the same road segment may be associated with a different station.

In addition, the numbers are not reliable. For instance, the VPH field contains very implausible values. Many times this value is more than 20 times the 15-minute count. For the purpose of the analysis, the raw 15-minute counts would be more applicable than the VPH field.

For each crash in the analysis period, a chart was generated showing the speed and the counts for the hour before and after a crash. In almost every case, it is possible to see a drop in speed or volume for the particular segment depending on if the crash occurred upstream or downstream of the detector. This implies that the counts are reasonably accurate assessment of the conditions at the time of the crash. Figure 3.3 shows two crashs' charts as an example.

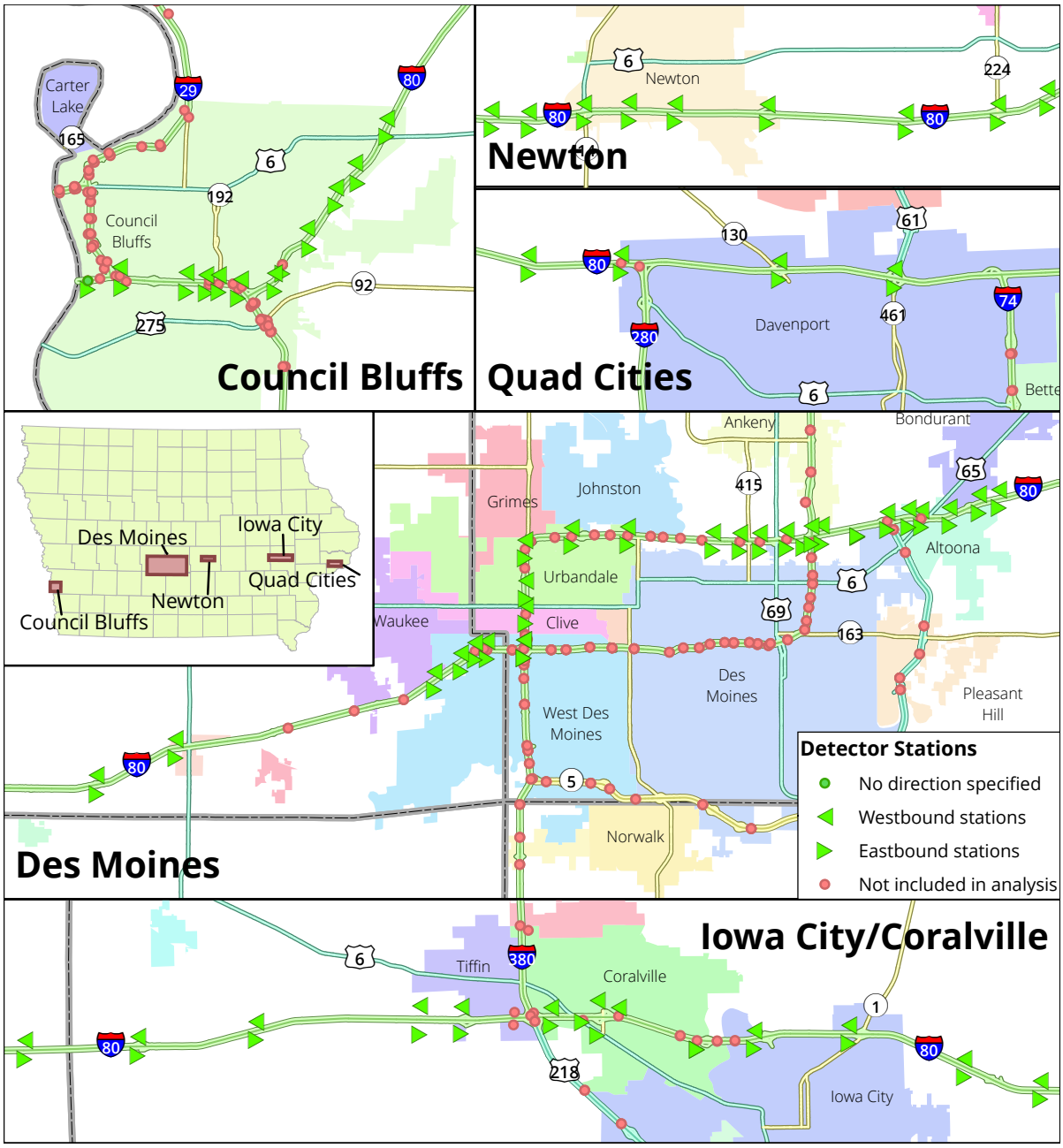


Figure 3.1: Map of stations in analysis

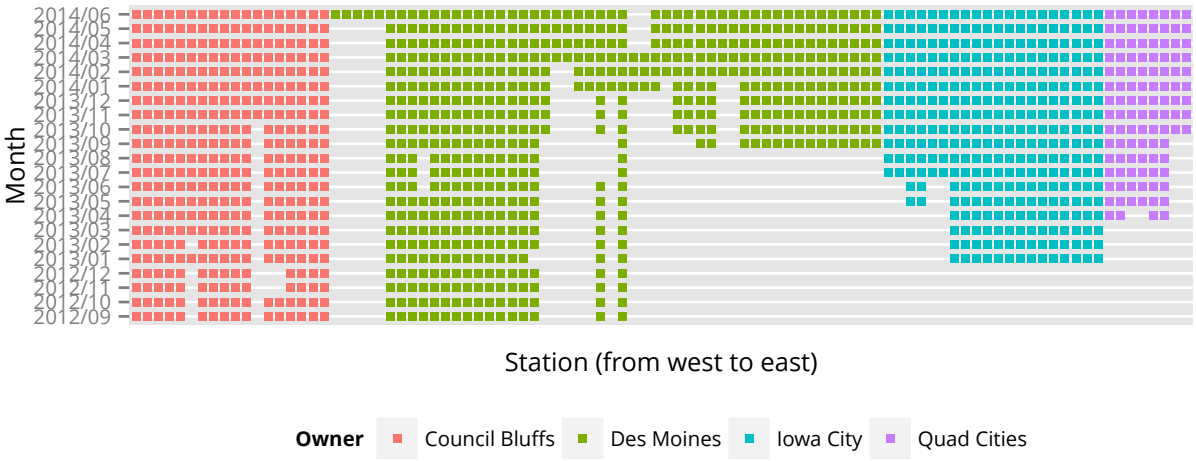
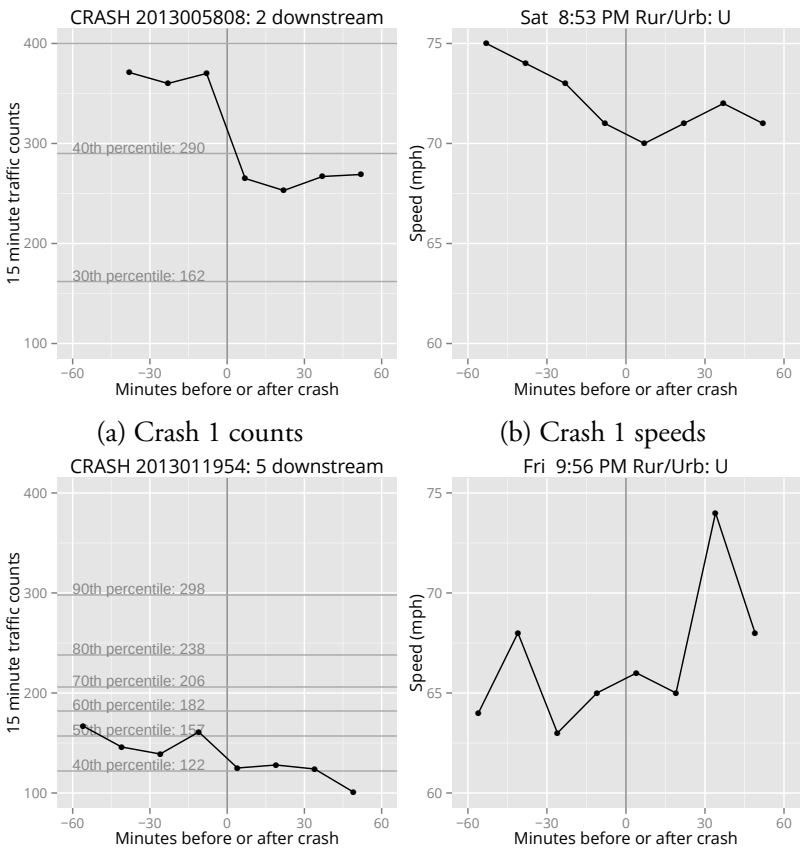


Figure 3.2: Availability of stations in analysis zone by date



(a) Crash 1 counts

(b) Crash 1 speeds

(c) Crash 2 counts

(d) Crash 2 speeds

Figure 3.3: Example speed and volume plots for two crashes in Wavetronix data

Additionally, there are many sections with traffic counts of zero. Some of these occur during high volume times and they should include at least one vehicle since there was an accident on the stretch of roadway. It is difficult to determine if these zeroes are caused by a lack of traffic or instrument malfunction. In the vast majority of records that have a count of 0, usually the occupancy and speed are missing, indicating issues with the instrument. The Wavetronix data generally follows expected speed/volume relationships when zero volumes and records with low quality values are excluded (an example is shown in figure 3.4; however, this is a significant portion.

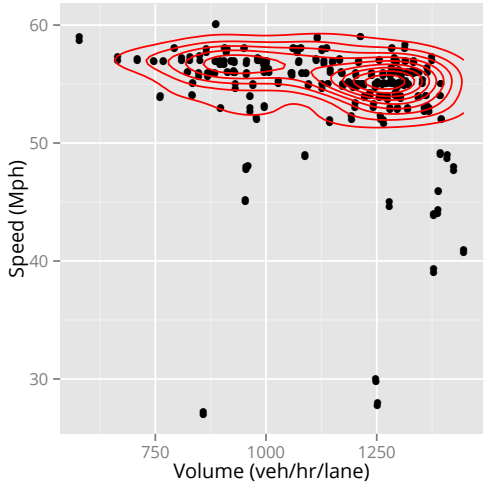


Figure 3.4: Speed vs. volume for the I-80/35 station @ Douglas avenue in Des Moines (August 2013)

Because of these issues, the Wavetronix data were not used in statistical analysis. There would be too few crashes near an operating station and the variability between different metro areas and lack of rural coverage would also hinder formulating a model. Instead, the Wavetronix data were used for comparing with other sources to confirm data and to explore the relationships between speed and volume along the I-80 corridor.

3.2 Automatic traffic recorders

The Wavetronix radar sensors are primarily used for real-time traffic managements and are therefore located in the larger urban areas in the state, leaving large gaps through rural areas. The Iowa DOT maintains automatic traffic recorder data for major roadways in the state. The primary

use for this is for traffic planning and to get aggregated traffic volumes along these major roadways, so it contains farther spaced stations covering a wider variety of regions when compared to the Wavetronix stations. For all but two stations that recorded to the quarter hour, the traffic counts are aggregated per hour. This time period is not extremely useful for determining the conditions at the time of a crash due to potential endogenity issues. An hour-long interval that includes a crash would likely have the counts affected by the crash, especially in severe cases where the crash reduces the capacity of the roadway.

However, these ATRs are very useful for calculating the average daily traffic [ADT] of a segment of roadway. As described in section 4.4.1.3, the sensors counts are aggregated together to determine the monthly ADT of each of the TMCs.

3.3 INRIX historical traffic speeds

INRIX is a commercial company that provides real-time traffic data throughout North America. Their data are mostly GPS data collected with in-vehicle transponders for commercial vehicles and increasingly with cell phones in passenger cars. The Iowa DOT has acquired historical traffic speed data for most major roadways in the state. Due to the vast amount of historical data, reasonably good estimates for traffic speeds when there are few vehicles reporting their speeds can be estimated using historical data. The INRIX data is available starting on January 1, 2011. The fields in the INRIX dataset are described in table 3.2

The INRIX data provides full coverage of the state of Iowa over the analysis period. When there are not enough vehicles to produce an accurate count, the historical averages for that segment are used to derive one. In the vast majority of these cases, the volumes are low and the speed is free-flow so the historical average is a good approximation. The INRIX data provide much more complete coverage of Iowa than the Wavetronix data both geographically and over time; however, it does not include volume data. Speed can be used as a proxy for congestion; low speeds generally indicate more congested roadways. Unfortunately, it is an imprecise relationship; it is especially difficult to distinguish between high-speed, low-volume and high-speed, moderate-volume conditions.

Table 3.2: Available fields in INRIX dataset

Field	Description
TMC Code	The Traffic Message Channel (see section 3.3.1) code for the reading
Measurement timestamp	The time of the measurement (to the nearest 5 minutes)
Speed	The estimated mean speed of the roadway segment (mph)
Reference speed	The calculated free-flow mean speed for the roadway segment (mph). Based on the 85th-percentile of all observed speeds
Historic average speed	The historic average speed for the same time of day and day of week (mph)
Confidence score	A simple score calculated by INRIX based on the confidence of the speed values: 30 = High confidence, real-time speed data used; 20 = Medium confidence, mix of real-time and expected speeds; 10 = Low confidence, primarily based on historical speed
C-value	Probability of reading representing actual roadway conditions, formula proprietary to INRIX. Only applicable when confidence value is 30

3.3.1 Traffic message channels

The INRIX data is segmented based on Traffic Message Channels. TMCs are used by commercial vehicle to deliver traffic and travel information using FM signals by encoding the information so that when there is an event, it can be broadcast to commercial vehicles and the TMC code can be used to indicate where the incident is occurring. These TMCs describe a roadway only based on landmarks such as mile-markers and exits. Table 3.3 contains statistics on the lengths of the TMCs.

Table 3.3: TMC length descriptive statistics (miles)

Minimum	25th percentile	Median	Mean	75th percentile	Maximum	Standard Deviation
0.015	0.55	0.74	1.75	2.84	8.41	1.79

3.4 Road weather information system

A Road Weather Information System [RWIS] is a system of sensors along the roadway that measure environmental conditions such as temperature, precipitation, wind and surface conditions. The Iowa DOT maintains historical readings from their RWIS system along major highways in the state dating all the way back to 1995. There are 14 stations located along I-80. Luckily, since weather conditions do not vary as much over long distances compared with other data such as speed and volume, these stations should be able to give reasonable estimations of the environment at the time of any crash in this time period.

Some of the major data collected by RWIS stations include:

- Air temperature (°F)
- Wind Speed/Gusts (knots)
- Wind Direction
- Pavement Sensor Temperature (°F)
- Pavement Sensor Condition (each station has 0–4 sensors)
- Subsurface temperature

The crash database also contains fields about the conditions at the time of the crash (see section 3.6). However, these data are categorical in nature and not nearly as fine grained as the RWIS data but they could prove useful to double check the accuracy of the RWIS data. This is explored in section 3.7

3.5 Geographic information management system

The Iowa DOT maintains a roadway information database, Geographic Information Management System [GIMS]. The roadway database contains multiple tables detailing the road database in a geospatial format including centerlines of all public roadways in Iowa. Major fields include road classification, name, access control, number of lanes, detailed lane information and AADT values.

However, there are some data that the GIMS dataset does not have. It does not account for road curvature or grade. The GIMS dataset is segmented so that GIS segments break at any intersection or when fields change. Note that for divided roads, there is only one centerline for both directions and not one for each direction.

3.6 Iowa DOT crash database

The Iowa DOT maintains public access to ten years of crash reports in a variety of tables. Each crash is geospatially located and there are many data tables available for each crash. There are three levels of data: crash-level (one record per crash), vehicle-level (one record per vehicle), and person-level (one record per person involved in each crash). Each person (aside from non-motorists) is associated with a vehicle using a unique key and each vehicle is associated with a crash using a unique key. These keys can be used to associate the same vehicle or crash across tables as well. Table 3.4 contains a list of all available tables and some of the notable fields in each table.

Each crash in the database is based off of the responding officers' report. The crash report format standard has been the same since 2001 using the fields described in table 3.4. After that, the responding agency forwards the crash report to the DOT which then processes them. Most squad cars in the state of Iowa are equipped with a GPS device to accurately locate the crashes. However, other crashes are manually located using a literal description of the crash location. Since the crashes in this analysis all occur on I-80, the crash locations should be relatively precise due to the abundance of mile markers and landmarks.

For the purposes of this analysis, crash data were collected from January 1, 2008 to June 30, 2014. The end date was chosen to maximize the number of crashes present in the analysis without including months that were missing data. The February 16, 2015 snapshot of the Iowa DOT crash database was used. Data for 2014 are considered preliminary due to reporting lag, but there were not expected to be many additions or modifications to crashes from the time period.

Table 3.4: Summary of available Iowa DOT crash database tables

Data level	Table	Notable fields
Crash level	Crash Point	X & Y coordinates, county, city
	Crash Type	First harmful event, manner of collision, major cause, drug & alcohol related
	Environmental	Weather conditions, light conditions, surface conditions
	Location and time	Date, time, roadway location, rural or urban, direction, overpass/underpass, crash location description
	Roadway	Route, vehicle direction, mainline/ramp, road classification, intersection class, roadway contributing circumstances
	Severity	Crash severity (KABCO scale), number of fatalities & injuries, property damage
	Work-zone related	Work-zone location, work-zone type, workers present
Vehicle level	Commercial vehicle	Number of axles, gross vehicle weight rating [GVWR], hazmat placard, hazardous materials released, license plate state
	Crash type	Driver sequence of events, most harmful event, fixed object struck
	Driver	Driver age, driver gender, driver charged, alcohol and drug tests, driver condition, driver contributing circumstances, vision obscurement
	Roadway	Speed limit, traffic controls
	Vehicle	Vehicle configuration, vehicle year, make and model, vehicle defect, initial direction, vehicle action
	Vehicle damage	Point of initial impact, most damaged areas, extent of damage, override/underride
Person level	Injured passengers	Injury status, gender, age, protection used, ejection, airbag, trapped, hospital transported to
	Non-motorists	Non-motorist type, location, action, condition, contributing circumstances
	Uninjured passengers	*Same fields as injured passengers with blanks for non-applicable data*

3.6.1 Crash data accuracy

This study assumed that the Iowa DOT crash data are accurate enough to not warrant manual correction aside from the direction of the crash and time that the crash occurred, both of which were manually inspected for every truck-involved crash in the time period. The crash reports have undergone through quality control by the DOT and should be accurate enough over the entire sample. The main goal of this analysis is to accurately get the time and place of each crash and use that to associate the crash with the real time traffic conditions. However, both the location and the time might be slightly inaccurate in the reports for a variety of reasons described below.

The accuracy of crash reports varies depending on the severity of the crash. Property damage and minor injury crashes—and single vehicle crashes in particular—often go unreported for monetary or other reasons. However, more severe crashes tend to be reported since there is almost always emergency personnel who need to respond to those incidents. Blincoe et al. (2002) estimated that 48% of PDO crashes, and 8–22% injury crashes are underreported, but that virtually 100% of critical or fatal injuries are reported.

The time is based on the police reports. They cannot be expected to be exactly accurate since the officer is rarely at the scene of the accident as it is occurring and must rely on witnesses to determine the time of the crash. The crash database only records crash times to the nearest fifteen minutes. The Wavetronix data is only in fifteen minute intervals, the INRIX data is in five minute intervals and the RWIS data precision varies by station and year but is generally accurate to the hour. For this study, it will be assumed that crash times are approximately accurate to the half hour and they may be adjusted based on the observed traffic conditions. Very few observations had to be adjusted.

The location can also be slightly inaccurate. For crashes where there is a police officer on the scene and the location should be highly accurate; these crashes are usually located via a GPS unit inside the squad car. Self-reported crashes depend on the description of the crash location provided

by the individuals filing the report, usually based on mile posts. This can be affected by whether the crash is rural or urban; urban areas have more frequent mile posts and landmarks so crashes can often be located more precisely.

3.7 Data validation

Since many data sources contain similar fields, it is easy to verify that all the data sources match up for each crash. The primary overlaps are weather data between the crash database and the INRIX speeds and the Wavetronix speeds. Table 3.5 shows a cross-tabulation of the police reported weather conditions and the conditions in RWIS for that day. Some days RWIS may record rain and snow and the police report allows for two weather conditions to be listed, so crashes might appear in multiple rows or columns. For the crash severity model, the surface condition from RWIS was used unless it had an error code, in which case the equivalent code from the crash report was used. "Slush" was considered the equivalent of "Ice and Snow".

Table 3.5: Cross tabulation of weather conditions in RWIS and the Iowa DOT crash database

	Police-reported surface conditions					
RWIS surface conditions	Dry	Wet	Ice	Snow	Slush	Other/Not Reported
Dry	831	45	12	8		79
Wet	34	130	12	23	12	3
Chemically Wet	3	4	34	27	3	2
Ice and Snow	44	15	150	154	9	2
Other/Error	182	38	36	32	2	28

Figure 3.5 shows a scatter plot of the average speed the hour before the crash for both INRIX and Wavetronix. Overall, there is a very clear correlation between the two. On average, INRIX speeds were slightly lower than Wavetronix data, but most Wavetronix records with speeds below 20mph did not have adequate quality control indicators. Overall, 80% of all readings are within

ten miles an hour between the two databases. For the purposes of this study, only INRIX data will be used and it will be assumed to be close enough to Wavetronix to not warrant any adjustment.

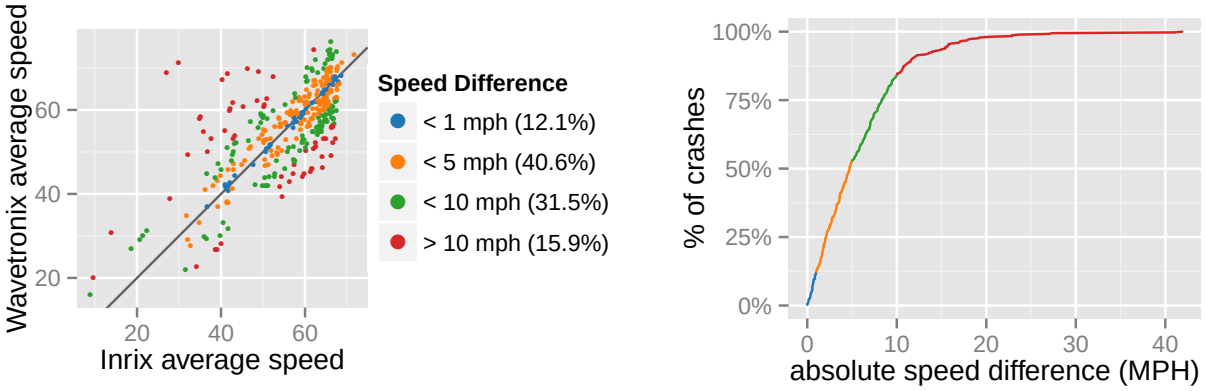


Figure 3.5: Comparison of speed data in INRIX and Wavetronix datasets

CHAPTER 4. METHODOLOGY

This chapter details the methodology for creating the models seen in the next chapter. The first step was to identify the extent of the study and which crashes would be included (described in section 4.1. After that came the most critical task, finding a way to associate all the different data sources together. Section 4.2 describes how a linear referencing system was used to locate all of the different stations, crashes and road segments and put them together. After this two statistical models were developed using the procedure outlined in section 4.4

4.1 Selected roadways and crashes

To analyze similar roadway segments and limit confounding factors from the roadway itself, a single route was chosen through Iowa. A freeway was needed in particular because the coverage of all the different data sources is more consistent on Iowa freeways. I-80 was the best candidate because it traverses across the state, passes through major metropolitan areas and has the most sensor coverage for Wavetronix data.

The entirety of I-80 through Iowa was chosen from the Missouri to the Mississippi Rivers—including concurrencies with I-29 and I-35—aside from one portion. At the western "mix-master" interchange in the Des Moines metropolitan area, vehicles wishing to go between I-80 and the I-80/I-35 concurrency requires traveling on turn ramps that only have one lane and do not conform to normal Interstate standards. This has not been included in the analysis as shown in figure 4.1. Other major concurrencies and interchanges allow traffic on I-80 to continue straight and on roads that are up to interstate standards and are thus included in the analysis.

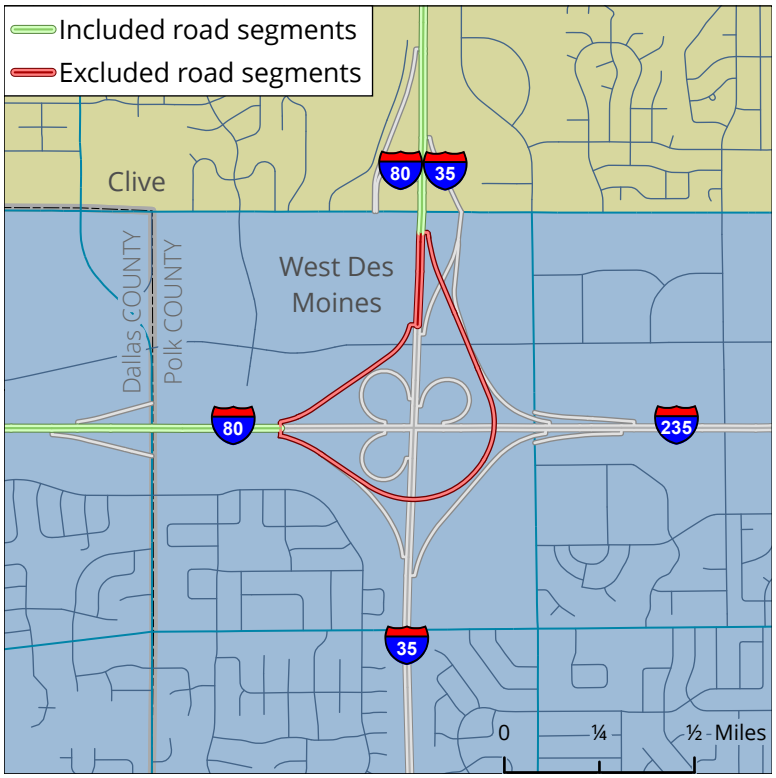


Figure 4.1: Included sections by the west mix-master interchange in Des Moines

4.1.1 Selecting crashes

This report seeks to analyze the effects of congestion on truck crashes in particular so two datasets of crashes were made. The first only includes crashes involving at least one truck and the second includes all crashes. The different vehicle classes that were determined to be a truck are listed in table 4.1.

For both datasets, the following queries were used to further refine the set of crashes down. The following criteria were used to limit the crashes to only include ones occurring along the I-80 mainline:

- *Road classification* = interstate
- *Ramp/mainline* = mainline
- *Route* = I-80, I-29 or I-35
- *Spatial proximity to I-80 linear referencing system* = < 300′ (see section 4.2.1)

Table 4.1: Included vehicle configurations

Truck vehicle configurations	Non-truck vehicle configurations
Single-unit truck (2-axle/6-tire)	Passenger car
Single-unit truck (3 or more axles)	Four-tire light truck (pick-up/panel)
Truck/trailer	Van or mini-van
Truck tractor (bobtail)	Sport utility vehicle
Tractor/semi-trailer	Motor home/recreational vehicle
Tractor/doubles	Motorcycle
Tractor/triples	Moped/All-Terrain Vehicle
Other heavy truck (cannot classify)	School bus (seats >15), small school bus (seats 9-15)
	Other bus (seats >15), other small bus (seats 9-15)
	Farm vehicle/equipment
	Maintenance/construction vehicle
	Train
	Other/not reported/unknown

Unfortunately, the crash database does not contain any information on I-80's concurrencies. Both of I-80's concurrencies (with I-29 in Council Bluffs and with I-35 in Des Moines) are coded in the crash database as occurring on I-29 and I-35, respectively. Therefore, all crashes occurring along I-80, I-29 or I-35 were included and then by inspecting the literal crash description and the location in GIS, crashes occurring on the concurrencies but coded to I-35 or I-29 were kept.

Each station is directional. In order to associate the crash with the correct direction , the "initial direction" variable (which is defined for each vehicle in the crash) was used to determine the direction of the crash. There were many different cases that required handling separately:

1. When all vehicles in the crash are heading either eastbound or westbound, that direction was chosen.
2. When all vehicles were either northbound or southbound (which occurred frequently during the concurrencies with I-80), the prevailing direction of the road segment was used to determine whether the crash occurred in the eastbound or westbound direction.
3. When the vehicle direction was unknown, the literal description was used alone with prefer-

ence for the direction the truck was facing for crossed centerline crashes. If no direction was given, then the crash was excluded from analysis.

For cases 2–4 above, the literal description of the crash location was used to supplement any other information. Most crashes contained which direction of I-80 the crash occurred on. Some sample literal crash descriptions include "NB/EB Interstate 0080 measuring 0.5 Miles West from (Milepost 256)" or "MM 127 EB/NB I-80/35." In these cases it was obvious what direction the crash occurred on and any vehicle directions were ignored. In addition, a search was made for "ramp" in the literal description. Any crash that occurred in the ramp yet was not coded as a ramp was excluded. Less than 1% of crashes had to be excluded.

4.2 Associating crashes with other data sources

With so many different data sources, a method to associate the crash with the different readings in the other data sources was needed. The linear referencing capabilities of ArcGIS provide a perfect way to measure the distances between crashes and stations.

4.2.1 Linear referencing system

A linear referencing system [LRS] was developed to correlate crashes with the different analysis datasets. A linear referencing system uses a contiguous linear route in GIS to locate other features as a distance along the route. This functions very similarly to the mile marker system on the Interstate system. Starting with mile 1 at the beginning of an interstate route in a particular state, all exits are numbered based on the miles from that starting point. Similarly, in this analysis, crashes, radar stations, roadway segments, and weather stations are located based on their distance along the route. This makes it easy to find the distance between two points simply by subtracting the mileages. These distances are as the driver sees them—as a distance along the roadway—instead of the straight line distances. Figure 4.2 shows a simplified diagram of an LRS.

4.3 Database design

All of the raw tables from each data source were added to a relational database: crash, vehicle and person–level crash data, Wavetronix measurements, INRIX measurements, and RWIS measurements. Each Wavetronix station, INRIX TMC and RWIS station was given a unique ID and located geospatially using its latitude and longitude coordinates. Then ArcGIS was used to find the mileage of each of these stations and the mileage was inserted in a new table in the database. To find the measured values (e.g. speed, volume, weather), the closest station/segment can be found by comparing the mileage. Then the measurement recorded at the time nearest to the crash at the closest station can be found. These relationships are shows in figure 4.3

4.4 Statistical analysis

Once all the data was aggregated as shown above, two different statistical models were developed. The first is a crash frequency model that analyzes the likelihood of a crash occurring on a given road segment within a specific time period. The second model is an crash severity model; it determines what factors correspond to higher severity crashes given that a crash occurred. These formulation of these two models is described in this section and the results are described in chapter 5.

Both models were formulated the same way. Since there are a large number of variables in the analysis, a base model with only a constant was created. Variables were added to this base model with the goal of maximizing the log-likelihood ratio of the models. Variables were only included in the analysis if they had a p-value that was below the significance threshold of 0.05. In addition, during the creation of models, the coefficients of the variables were tracked. Some of the fields in the analysis are highly correlated, so there is a possibility of multicollinearity which often causes dramatic swings in parameter estimates. When parameter estimates changed dramatically, the correlation between the independent variables was checked to ensure that there was not any correlated dependent variables.

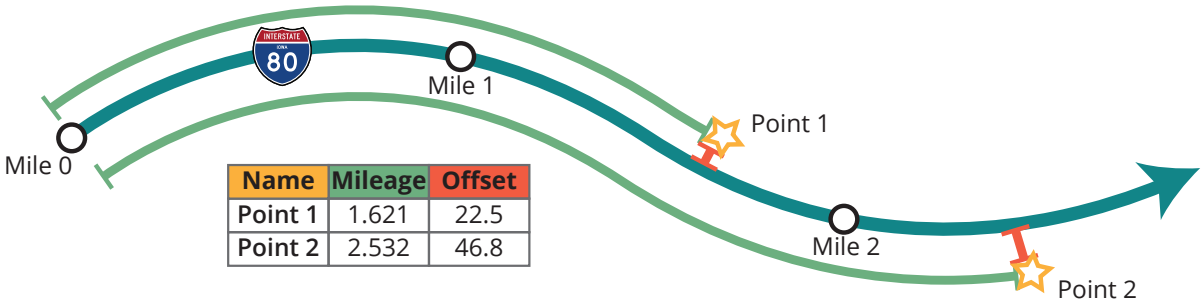


Figure 4.2: Explanation of a linear referencing system. Geospatial points (yellow) are located along an LRS route (blue). The distance along the route is recorded as the mileage (green) and the distance from the route is recorded as the offset (red).

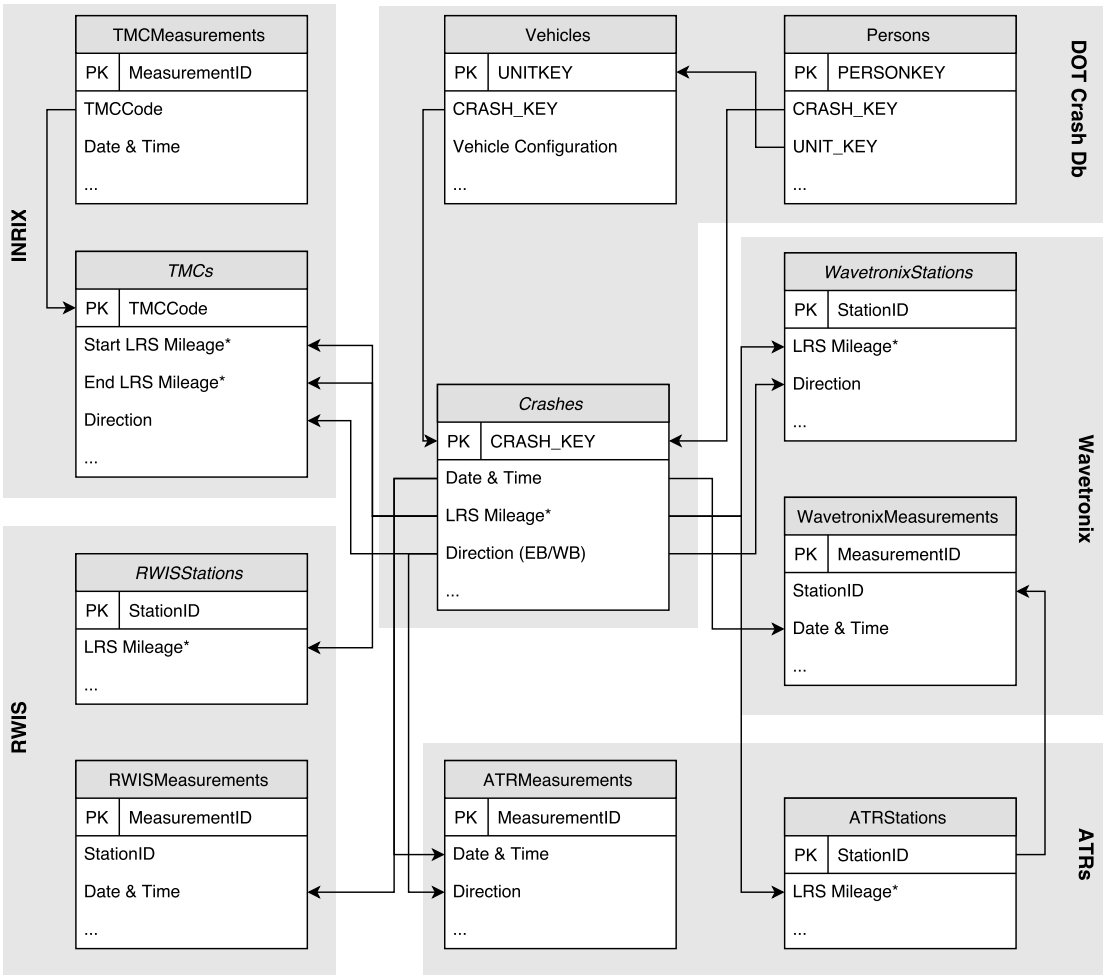


Figure 4.3: Diagram of relationships in database

The final models were compared to the base case using a χ^2 test (section 2.3). In the case of the negative binomial model, it was also checked against the Poisson distribution with the same parameters to check for overdispersion.

4.4.1 Crash count model

The first model created was a model to determine what factors cause more crashes occurring on a given segment in a given time frame. For this, a random parameter negative binomial model [RPNBL] was used. This is discussed in more detail in section 2.4. For this analysis, each record represents a single road segment over a given month. The dependent variable was the total sum of all crashes on that specific segment in a given month. The road segments used were the TMCs used by INRIX. They have a generally uniform size and characteristics and offered significant enough length to have multiple crashes. NLOGIT 5 was used to estimate the crash count models.

After considering various model forms, it was found that a random-parameter Poisson [RPP] is more appropriate than an RPNBL. Even though the variance of the dependent variable is approximately three times its mean, the RPNBL failed a couple diagnostic tests: the dispersion parameter α was not significant. In addition, the RPNBL model failed the likelihood ratio test when compared to the equivalent RPP model.

When formulating the random parameter model, each variable was initially formulated as a randomly varying coefficient, assuming to have a normal distribution. The coefficients were set to vary randomly across TMC road segments. If the coefficient for the standard deviation of the parameter was not significantly different from 0 then the coefficient was assumed to be fixed.

The final model is as follows, where y is the number of crashes on a given segment in a month, X is the matrix of predictor variables and β_{ij} is the vector of coefficients for observation j on TMC segment i. Since this is a random parameter model, the coefficients β_{ij} for the randomly-varying parameters are allowed to vary by TMC according to a statistical distribution (in this case, a normal distribution). The standard deviation, σ_i of the random coefficients is estimated by the model as well as the individual coefficient means, β_j

$$log(\mathrm{E}[y|X]) = \sum \beta_{ij} X_{ij}$$

$$\beta ij = \beta + \varphi_i$$

$$\varphi_i \sim N(0, \sigma_i)$$

By rearranging this equation, it is possible to get the expected number of crashes directly using the equation $\mathrm{E}[y|X] = e^{\beta_i x_i}$

4.4.1.1 Roadway characteristics

The INRIX TMCs are not coincident with the GIMS dataset, which contains roadway data such as speed limit, shoulder width, number of lanes, etc. In general, the TMCs had uniform characteristics, which facilitated manually determining the characteristics for the roadways. The method of determining the characteristics for each targeted characteristic is described below.

- **Number of lanes:** The vast majority of I-80 has two lanes in each direction in Iowa aside from areas of the Council Bluffs, Des Moines and Iowa City/Coralville metropolitan areas which have three lanes in each direction. The number of lanes variable was manually populated using visual inspection in ArcGIS.
- **Median type:** The median type also corresponded well to the TMCs. There were four median types in the study and they were assigned to TMCs using visual inspection in ArcGIS
- **Shoulder width:** The shoulder width was calculated via a weighted average from the underlying GIMS segments.
- **Lane width:** The lane width in GIMS for almost all of I-80 is 12′. It was not used in this study.
- **Rumble strips (left and right):** The rumble strip was present on the left and the right for the entire corridor and was not used.
- **Surface Type:** The surface type for the entire roadway fell into two major categories, hot-mix asphalt [HMA] and Portland cement concrete [PCC]. A variable was made called Per-

Table 4.2: Fields aggregated from INRIX

Fields	Description	Thresholds
Percentiles	The Xth percentile speed recorded over a month	0.001, 0.005, 0.01, 0.05, 1, 2, 3, 5, 10th percentiles
Percent of speed limit	The percent of time spent below XX% of the speed limit	10, 20, 30, 40, 50, 75, 85, 95, 100, 105 and 110%
Absolute difference from speed limit	The percent of time where the prevailing traffic conditions are XX mph lower than the speed limit	15, 10, 5, 1, 0

centHMA which is the percent of the TMC length that was paved with HMA. The percent paved with PCC would be 1 – PercentHMA.

Each of these variables was spot checked using aerial imagery and the records in GIMS were consistent with what was observed in aerial imagery.

4.4.1.2 Aggregating speed data

For each month, the INRIX dataset was used to determine what the prevailing traffic conditions were over the entire month. For each month, the following statistics were generated for the month. All records were kept, even ones with a low confidence value. This was done since the records with low confidence values generally happen more in rural areas and during the late evening hours. It is expected that during these times, traffic would be low and speeds would be near free-flow.

Microsoft SQL server was used to preform this aggregation. The aggregated values are listed in table 4.2.

4.4.1.3 Calculating vehicle miles traveled

Commonly when doing crash frequency models, a variable indicating exposure is needed. This is because the incidence of a crash is a probabilistic even that depends on how much travel is done. The more vehicles traveling, the more likely that a crash will occur. Typically, the measurement used

is vehicle–miles traveled [VMT]. For Poisson and negative binomial models, the log of the VMT is a very useful variable since the dependent variable is logged. If the coefficient of the exposure variable is fixed to be 1, then doubling the VMT will lead to doubling the number of predicted crashes. As shown in section 5.1, the estimated coefficient of the log of the VMT was not significantly different from 1, which validates that this is a good method of measuring exposure. Including AADT and length as separate exposure variables was also explored but provided worse fit.

The GIMS dataset provides AADT data for every year up to 2013 for the entire Iowa DOT roadway network. Most of this is derived from ATR data (see section 3.2 and other portable traffic counters. Unfortunately, these counts are not broken down by month. There are large seasonal volume variation. To get more accurate monthly traffic volumes, a set of adjustment factors were created for each ATR station along I-80 by calculating the ratio of the monthly ADT with that year's AADT. These factors are included in Appendix B Every GIMS record's AADT was multiplied by the factor for the specific month and year of the ATR station closest to it.

- Most stations have complete records over the whole time period. For these records, the factor is just the monthly ADT divided by the AADT.
- In other cases, the average factor for that specific time period over the other three years is used.
- When the AADT is missing from a particular year for a station due to missing records, the AADT from GIMS was used.
- For 2014, GIMS does not provide any AADT data for the roadway segments. To estimate the AADT for segments, the 2013 AADT is multiplied by the ratio of the AADT between 2013 and 2014 for the nearest ATR station.

From that point, the total VMT for each TMC in the course of a month can be calculated by multiplying the AADT, the length of each GIMS overlap with the TMC and the number of days in that specific month.

4.4.2 Ordered Severity Model

The second model determines, given that a crash has occurred, what factors are associated with severe crashes. For this model, each row represented a single crash from the Iowa DOT crash database (section 3.6). Due to the low number of fatal and major injury crashes compared to other severity models, for the ordered probit, the categories were aggregated into three groups: fatal and major injury crashes, minor and possibly injury crashes; and property damage only. The models were estimated in R using the "polr" package.

Some of the Iowa DOT crash database had fields with null values. For any variables that involved null values, indicator variables were used, ensuring that null cases were not excluded for the model. For instance, the two axle-related variables were indicator variables for having either ≤ 4 or ≥ 7 axles. For both variables, null values were treated as being false.

There are two main ways to formulate an ordered probit model are to either estimate $n - 1$ threshold values or to estimate $n - 2$ threshold values and a constant term (where n is the number of discrete categories). This model uses the former. Therefore the model is formulated as follows where y_i^* is the latent continuous variable, x_i are the predictor variables for observation i, β are the estimated coefficients, and ε_i is a normally distributed error term:

$$y_i^* = \beta x_i + \varepsilon_i$$

The predicted crash severity can then be found by comparing the y_i value to the estimated thresholds μ_1, μ_2.

$$y = \begin{cases} PDO \text{ if } & y_i^* \leq \mu_1 \\ B/C \text{ if } & \mu_1 < y_i^* \leq \mu_2 \\ K/A \text{ if } & \mu_2 \leq y_i^* \end{cases}$$

4.4.2.1 Traffic conditions

The roadway conditions from the time of the crash were gathered from the INRIX dataset by grabbing the hour before and after the crash. This was placed in a "wide" format for analysis, where each five-minute reading in the two-hour period was placed in its own field. Only the speed and confidence values were kept. For the analysis, the speeds for the hour after the crash were not used; these values are endogenous because a more severe crash is more likely to affect the traffic stream when compared to a minor fender bender.

From these variables, many more were calculated. For each of these, multiple analysis periods were chosen: the full hour, the half hour and the quarter hour preceding the crash.

- The maximum, mean and minimum recorded speeds preceding the crash
- The variance of the speeds preceding the crash
- The absolute and relative difference between speeds prior to the crash (e.g. the difference between the speed 5 minutes before the crash and the speed 15 minutes before the crash)

CHAPTER 5. RESULTS

This chapter contains the results of the regression models, findings and and a discussion of the findins and the models' goodness of fit measurements. Summary statistics for all of the derived parameters included in the model are shown in appendix A.

For all of the results below, the superscript after the variable indicates which data source the variable came from. The following data sources were used in the models:

C The Iowa DOT Crash-level database (section 3.6)

V The Iowa DOT Vehicle-level database (section 3.6)

I INRIX Speed data (section 3.3)

G GIMS (section 3.5)

A ATR traffic recorders (section 3.2)

R RWIS (section 3.4)

5.1 Crash frequency model

Table 5.1 contains the results of the crash frequency model. The formulation of this model was described in greater detail in section 4.4.1. The final model includes five fixed parameters and four random parameters. All parameters are significant at a level of 0.05. The constant in the crash frequency model is highly negative; this matches the data because the majority of all segments have no crashes. All percentage variables are on a scale from 0–100 and the random parameters vary by TMC.

Table 5.1: Results of crash frequency model

	estimate	**std error**	**t-value**	**p-value**
Non-random parameters				
Constant	−36.83	1.328	−27.71	$< 10^{-16}$
Percent of time speed was above speed limit[I]	−0.007	0.002	−3.97	$3.56{\cdot}10^{-5}$
log(Monthly VMT) in hundred million VMT[AG]	0.992	0.035	27.96	$< 10^{-16}$
Random parameters				
log(Percent trucks)[G]	0.244	0.108	2.25	0.027
σ	0.084	0.008	9.78	$< 10^{-16}$
Left shoulder width (ft)[G]	−0.063	0.019	−4.41	$1.1{\cdot}10^{-5}$
σ	0.024	0.004	6.17	0.0003
Percent of time in month with icy conditions[R]	1.736	0.236	7.34	$1.09{\cdot}10^{-13}$
σ	1.17	0.157	7.43	$5.4{\cdot}10^{-14}$
Percent of time with speed slower than 10mph below limit[I]	0.083	0.016	4.99	$3.1{\cdot}10^{-7}$
σ	0.079	0.086	5.94	$1.3{\cdot}10^{-9}$
Month is January or December (indicator)	−0.473	0.084	−5.64	$8.6{\cdot}10^{-9}$
σ	0.577	0.062	9.29	$< 10^{-16}$
Log-likelihood	$-3,517.7$			
Restricted log-likelihood (non-random model)	$-3,704.5$			
Restricted log-likelihood (constant only model)	$-4,031.6$			
McFadden's Adjusted R^2 (non-random model)	0.050			
McFadden's Adjusted R^2 (constant only model)	0.116			
Likelihood ratio χ^2 test statistic & significance (non-random model)	370.8	$< 10^{-16}$		
Likelihood ratio χ^2 test statistic & significance (constant only model)	933.5	$< 10^{-16}$		

5.1.1 Speed related variables

The first variable, the percent of the time in the month that the speed was above the speed limit, has a negative coefficient. The more time traffic is in free-flow speeds, the less likely there is to be a crash. This is intuitive because when there is less congestion, vehicles are more able to maneuver in the traffic stream and avoid other vehicles and objects. This variable has a slight negative skew as shown in figure 5.1, but transforming the variable did not improve fit.

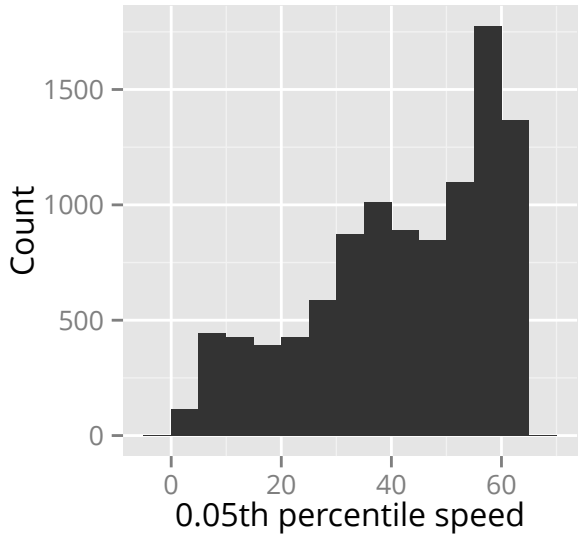


Figure 5.1: Distribution of the 0.05th percentile speeds

The other INRIX variable is the percent of time with the speeds slower than 10 miles per hour under the speed limit. This has a negative coefficient which reinforces the findings before that a decrease in speed leads to an increase in crashes. This is a random parameter. For 14.07% of the TMCs in the dataset, this coefficient is positive. The random parameter can be explained because different recorded speeds imply different traffic stream characteristics (e.g. the flow when speeds are around 5mph is much different than the flow when traffic is 30mph) and different road segments are more likely to have different safety characteristics that are affected by congestion in different ways.

Both higher speeds (the percent of time speeds were above the speed limit), which usually correspond with free-flowing conditions, and lower speeds (the percent of time where speeds were

slower than 10 mph below the speed limit), which correspond with congested conditions, were analyzed in the model. The same conclusion can be drawn from these variables: slower speeds in a given month are correlated with more frequent crashes.

5.1.2 Roadway related variables

The major roadway exposure variable is the log of the VMT. The coefficient is not significantly different from 1. As described in section 4.4.1.3, the way that the Poisson regression is constructed, a variable that is logged and that has a coefficient of 1 is linearly related to the dependent variable. The percent trucks variable has a positive coefficient. This is intuitive; if there are more trucks on the road, there are likely to be more crashes involving trucks.

In the formulation of the model, many exposure variables were tried. A combination of AADT and length was explored but the combination of the two variables was not as significant as VMT alone. There were no significant interaction effects between the AADT and the length variables. Also, truck VMT was explored. However, the truck VMT did not provide as significant fit and did not account for any increased exposure from non-truck vehicles. Since the monthly VMT was derived from the nearest ATR station, having a coefficient of 1 indicates that the factors applied reflect actual month-to-month variation. The mix of percent truck and VMT provided the best statistical fit but still took into account both the overall traffic and the amount of truck traffic on the roadway.

Another roadway characteristic, the shoulder with, indicates that the wider the left shoulder is, the fewer crashes there are. In almost all cases, the right shoulder along I-80 is within a couple feet of 12 feet wide. The left (inside) shoulder width varies significantly over the length of I-80. When it is wider, vehicles have more of an opportunity to correct back into the correct lane when they stray over it. Trucks may be particularly sensitive to shoulder width; they are wider than passenger cars and require more space to maneuver to correct potential roadway departures. This variable is random, but the coefficient is almost always negative. In general, shoulders were wider in urban areas (the indicatore variable for urban areas was not significant).

5.1.3 Time and weather variables

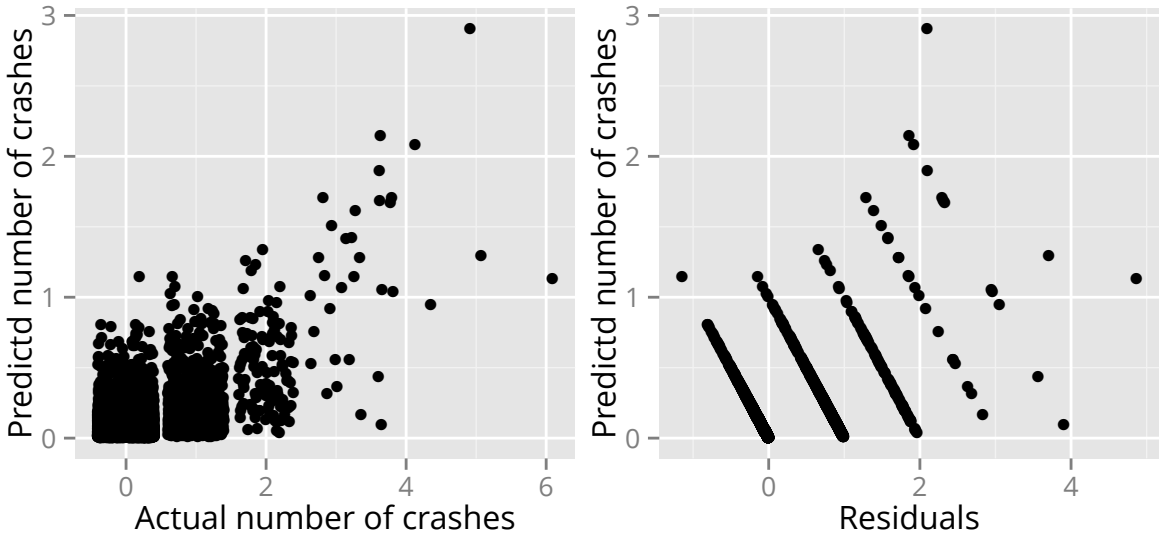
The variable for percent of time in the month with icy conditions is positive. Snow and ice on the roadway lead to an increase in crashes by making the task of driving and maintaining control much more difficult. Since trucks are already more difficult to control than passenger cars, hazardous weather can be especially tricky for trucks to navigate.

The coefficient for records in December and January is negative; this is similar to other studys' findings showing summer months as the most crash-prone. The interaction between the month and the percent of time icy was not significant. When conditions are dry and clear, there are fewer crashes in these months than others. Likely causes include fewer recreational trips and less drinking and driving in the cold weather months. However, there are many other factors that can affect a months' crash rate including any special events or construction. This is captured by the random parameter; the coefficient for January and December have a larger standard deviation than the other random parameter values, the standard deviation exceeds the mean. This indicates that different TMCs did not have similar crash patterns in December and January. Similar indicator variables for other times of the year were not significant.

5.1.4 Model diagnostic

The likelihood ratio test was the major diagnostic tool to determine the suitability of the final model. The model was compared to a Poisson model containing only a constant and a non-random Poisson model. Simply comparing the log-likelihood ratios of each of the models shows that the random-parameter Poisson model has the closest fit. Since the additional parameters in the model alone would increase the fit, the likelihood ratio test (described in section 2.3) was used to compare the random- parameter Poisson model to a non-random Poisson regression and to a model using only a constant. In both cases, the likelihood ratio test statistic was very significant, indicating that the random-parameter Poisson model describes the variability much better than the alternative models. An alternative random-parameter negative binomial model was also explored; however, this model did not pass the likelihood ratio test.

One potential downfall of this model is the low mean. The majority of road section–months in this study have zero crashes. The mean number of crashes per month per segment was 0.121. Figure 5.2 shows the predicted value and residual plots. Figure 5.2a shows the predicted number of crashes on the y axis and compares it to the actual number of crashes on the segment (jitter has been applied to the points in the x direction, all actual number of crashes are positive integers). Figure 5.2b shows residuals (the difference between the predicted number of crashes and the actual number of crashes) compared to the predicted number of crashes. These plots indicate some issues with the model specification. Mainly, the assumption of homoskedasticity (the assumption that the variance of the residuals is constant) is violated in the residual plot because the spread of points increases as the predicted y values get larger.



(a) Predicted vs. actual crashes

(b) Residuals vs. predicted value

Figure 5.2: Predicted values and residuals for the crash frequency model

As discussed in section 2.3 on page 9, Wood (2002) developed a goodness of fit measure for when the dependent variable has a low mean. For this dataset, the calculated test statistic is 23.6. This is extremely significant ($< 10^{-16}$), indicating that the model may not provide a good fit. However, the authors of the paper admit that with a large sample size like this one, it is likely that even a good model will pass the test due to the large number of degrees of freedom used.

Table 5.2 contains the predicted number of crashes (by rounding $\hat{\mu}_i$ to the nearest integer value) compared to the actual number of crashes. Italicized cells indicate correct predictions (90.0% of records are predicted correctly when rounded to the nearest whole number) Since the mean number of crashes on a given segment in a month is very low, the Poisson model underpredicts the number of crashes for months that have a large number of crashes. This limits the predictive power of the model, but useful inferences can still be obtained from the magnitude and the direction of the coefficients in this model.

Table 5.2: Crash frequency model predictions

Predicted	**Actual # Crashes**						
# Crashes	0	1	2	3	4	5	6
0–0.5	*9,141*	871	82	12	5	0	0
0.5–1.5	63	*46*	24	7	6	2	1

Summaries of the random parameters are listed in table 5.3. The first column is the name of each variable. The estimate column has the mean coefficient estimate of each variable. The second column is the standard error of the coefficient estimate. The estimate divided by the standard error results in a t-statistic. The t-statistic can produce a p-value for a standard student's t distribution to determine the two-tailed probability of the null hypothesis of the coefficient equaling zero. For the random parameters, a second parameter, σ shows the estimated coefficient of the standard deviation of the normal distribution for the variable. The estimated coefficient mean and standard distribution can be used to determine the distribution of the variable across the sample population.

Since each of the coefficients are randomly distributed, there is a probability that the sign for a specific coefficient for a specific record will differ in sign from the coefficient mean. The probability of that is listed in the "sign change" column. The last two columns list the upper and lower bounds that 95% of records will fall into. For the log(Percent trucks) and left shoulder width variables, the 95% range does not include 0, implying that the majority of records will have coefficients of the same sign.

Table 5.3: Summary of random parameter distribution

Parameter	Mean	Std. dev.	Sign change	95% bounds lower	95% bounds upper
log(Percent trucks)	0.2440	0.0843	0.19%	0.0788	0.4092
Left shoulder width	−0.0633	0.0241	0.43%	−0.0161	−0.1105
Percent time icy	1.7361	1.1720	6.93%	−0.5610	4.0332
Percent time speed > 10mph below	−0.0831	0.0792	14.70%	−0.2383	0.0721
January/December	−0.4735	0.5766	20.57%	−1.6036	0.6566

5.2 Crash severity

The results of the crash severity model are listed in table 5.4 on page 49. The processes used to make this model are described in section 4.4.2. The final model contains 16 estimated coefficients in addition to two estimated cutoff points. All variables in the final model were significant at a level of 0.05.

Only one variable from the INRIX data set was significant in the final model. The first variable captures the mean of each of the 5-minute periods before the crash. As discussed on page 21, the speeds with low confidence values can generally be trusted because they usually indicate free-flow conditions, so they are included in this mean. High-confidence speed data are available for 95.4% of crashes. This variable has a positive coefficient, which is backed up by previous literature showing that higher speeds generally lead to more severe crashes since faster vehicles are more difficult to control and have higher energy impacts. Likewise, congested situations are more likely to have fender-bender crashes which often do not has as significant of injuries. However, truck-involved crashes may be more likely to have injuries due to the larger size of trucks. This is captured by some of the other variables such as the presence of multiple trucks and the presence of non-truck vehicles.

The coefficient for crashes where at least one driver tested positive or refused a test for drugs and alcohol is positive; this is expected since drugs and alcohol impair driver ability and often lead to fatalities and incapacitating injuries. Weekends tend to have lower severity crashes. There is likely

Table 5.4: Results of crash severity model

	estimate	**std error**	**t-value**	**p-value**
Mean speed 30 minutes prior to crash	0.01345	0.00675	1.99	0.0462
Winter weather (1="Chemically wet", "icy," or "snowy", 0=other)	−0.28524	0.14426	−1.98	0.0480
Drug or alcohol related crash (1=at least one driver tested positive or refused drug/alcohol test, 0=otherwise)	1.32637	0.34167	3.88	0.0001
Weekend (1=Saturday or Sunday, 0=otherwise)	−0.30244	0.12759	−2.37	0.0179
Major cause (1=Swerving/evasive action, 0=other)	0.76988	0.14474	5.32	$1.17 \cdot 10^{-7}$
Major cause (1=ran off road:right, left or straight; 0=other)	0.87304	0.13764	6.34	$2.82 \cdot 10^{-10}$
Manner of collision (1=sideswipe, 0=other)	−1.01868	0.13790	−7.39	$2.20 \cdot 10^{-13}$
Multiple trucks present (1=multiple trucks, 0=one truck present)	0.92658	0.16287	5.69	$1.48 \cdot 10^{-8}$
Multiple non-truck vehicles present (1=non-trucks present, 0=only trucks)	0.67086	0.15465	4.34	$1.52 \cdot 10^{-5}$
A truck rear ended a vehicle (1=at least one truck point of initial contact is back of vehicle, 0=otherwise)	0.36389	0.15393	2.36	0.0181
Threshold between PDO and B/C Crashes	2.388	0.486	5.108	$1.79 \cdot 10^{-7}$
Threshold between B/C and K/A Crashes	4.684	0.485	9.661	$< 10^{-16}$
Log-likelihood	$-1,142.3$			
Restricted log-likelihood	$-1,228.5$			
McFadden's adjusted R^2	0.062			
χ^2 test statistic and p-value	172.4	$< 10^{-16}$		

to be some minor correlation between impaired driving and weekend; however, the correlation is only 0.075

Crashes caused due to swerving are more likely to be severe than other crashes. Swerving is a particular concern for semi trucks due to their higher profile and decreased maneuverability. Run-off-road crashes tend to have a higher crash severity. Many of these crashes have a car hitting a fixed object off the road at full speed, which is a particularly dangerous situation. Sideswipe crashes are less likely to be severe; sideswipe crashes often only involve property damage since it is a collision that generally involves vehicles travelling at similar speeds in the same direction.

When the pavement is covered in snow, ice or de-icing chemicals, crashes tended to be more severe. While speeds tend to be lower during poor weather, maintaining control of vehicles gets much more difficult and drivers are less able to avoid striking objects and are more likely to spin out and be impacted at a more direct angle.

The variable for more than one truck is positive. This makes sense since there are more opportunities for drivers to be injured and multiple trucks implies a collision between the trucks. Similarly, if a non-truck vehicle is present there is more likely to be an injury since a vehicle is much smaller than a truck and more likely to have multiple passengers who may get injured. The indicator variable for when a truck rear-ends a vehicle is positive. This is the same as the result found in Duncan et al. (1998).

5.2.1 Model diagnostic

Overall, the model does not violate any major assumptions. Unfortunately, it under-predicts the occurrence of fatal and major injury crashes due to the uneven distribution of severities. The same model without grouping the different severity levels does have a better log-likelihood but has fewer significant variables. Also, this model does not take into account a significant factor for crash reports: underreporting. As discussed by (Savolainen et al., 2011), less severe crashes often go unreported to avoid fines or changes to automobile insurance. While the coefficients in this model are useful to determine what helps determine crash severity, there is not much predictive power in

Table 5.5: Ordered severity model predictions

Predicted	**Actual severity**				
Severity	K	A	B	C	O
K/A	*0*	*0*	0	0	0
B/C	3	3	*15*	*15*	19
O	19	40	139	219	*1,423*

the model.

Table 5.5 contains the predicted severity group compared to the actual severity. Italicized cells indicate a correct prediction. In total, 1,453 of records of 1,895 (76.7%) were predicted correctly. The number of major and fatal crashes were severely underpredicted. This is common in ordered discrete models with highly unequal groupings. Even still, the directions, magnitudes and marginal effects of the variables provide useful insights.

Table 5.6 contains the marginal effects of the ordered severity model centered on the means of the independent variables. The marginal effects shows how a change in a variable affects the probability of a specific outcome occurring. For continuous variables, the marginal effect shows the change in probability in a specific outcome when the value is increased by 1. For an indicator variable, the marginal effect shows the change in probability of a specific outcome as the values change from 0 to 1

Table 5.6: Marginal effects of ordered severity model

	O	**BC**	**KA**
Mean speed 30 minutes prior to crash	−0.002	0.002	0.000
Winter weather	0.044	−0.037	−0.007
Drug/Alcohol-related	−0.293	0.227	0.066
Weekend	0.054	−0.045	−0.008
Major cause: Swerving	−0.149	0.123	0.026
Major cause: Ran off road	−0.167	0.138	0.029
Manner of collision: sideswipe	0.162	−0.137	−0.024
Multiple trucks present	−0.186	0.152	0.034
Multiple non-truck vehicles present	−0.112	0.095	0.017
A truck rear ended a vehicle	−0.066	0.055	0.011

CHAPTER 6. CONCLUSION

The INRIX and Wavetronix datasets up to this point have been mostly used for incident management and operations but there is a significant safety application that is waiting to be explored. The ability to get an accurate representation of flows and speed at any time in recent history and recent increases in computer storage capacity and power have increased the amount of data that can be stored and processed. Similarly, truck crashes are a relatively significant problem in traffic safety but there is still a lack of research demonstrating how these crashes differ from crashes involving other vehicles. Improved statistical methods and estimations increasingly allow for more accurate and useful models.

This thesis developed a framework to associate crashes with a variety of different data sources of a bunch of different geospatial types: INRIX road segments, sparse RWIS stations, dense Wavetronix stations and crashes. Using a linear referencing system allowed all of these to be associated together in a robust manner. This was simplified by selecting only one highway corridor, but the same solution could be easily applied to a more comprehensive network.

6.1 Major findings

The two models in this paper work together to get an idea of how frequent and severe traffic crashes on I-80 are. The Poisson model is the first tier which determines the likelihood of a crash or multiple truck-involved crashes occurring in a given month. The ordered probit model then determines, given that a crash occurred, how likely it is to have different injury severities.

The crash frequency model predicts the number of crashes on a given road segment in a given month. The major findings for the crash frequency model include:

- in general, months that had a lower speed tended to have more crashes. This is determined in the model by looking at the percent of time where speeds are slower than 10mph below the speed limit and the percent of time spent above the speed limit.
- The number of crashes is almost exactly linearly related to the VMT. If the traffic was doubled in a given month, the expected number of crashes would be doubled as well. Additionally, the number of truck crashes is affected by the percent of trucks on the road. Increase the percent of trucks and more truck-involved crashes would be expected.
- Icy conditions increase the number of expected crashes while wider shoulders are associated with a lower crash risk. All other things being equal (including weather), January and December have statistically significant lower risk of crashes than other months.
- Some of the dependent variables had statistically significant randomness, meaning the magnitude of their effects varied depending on the TMC. This helps to capture the range of uncertainty of the estimates as well as account for unexplained factors.
- The model tends to underpredict months that have a large number of crashes. The mean number of crashes has a low mean and there's evidence that there may be some bias in the model.

The major findings for the crash severity model include:

- Higher speeds preceding a crash are linked to higher crash severity. Crash severity also increases if the speeds are increasing over the time period.
- When more vehicles are involved in a crash, it leads to more severe crashes. Multiple trucks and multiple vehicles in the crash are more likely to have more crashes.
- Sideswipe crashes tend to be lower severity while crashes caused by swerving, run-off-road crashes and crashes involving drugs and alcohol tend to be more severe.
- Crashes are expected to be more severe when the pavement is snowy or wet. Weekend crashes were expected to be more sever.e

6.2 Further research

In the future, the data provided by both INRIX and Wavetronix will be much more thorough and suited for analysis. The limited time that these data sources have been available limit the number of crashes that can be associated with them which makes it difficult to get an adequate sample size. As time goes on, there will be more time periods where all the different data sources overlap.

In addition, the stations will become more reliable and detailed in the future. The INRIX dataset will soon include XD segments, which are much more granular than the TMCs currently used and will include approximate flow data. Wavetronix will become more pervasive as new stations are installed and the old ones experience more uptime.

APPENDIX A. DESCRIPTIVE STATISTICS

Table A.1: Crash frequency model descriptive statistics (continuous variables)

Variable	Mean	Std. dev.	Min	Max
Number of crashes	0.121	0.398	0	6
Number of fatal crashes	0.001	0.036	0	1
Number of injury crashes	0.028	0.180	0	3
Number of PDO crashes	0.091	0.333	0	5
Speed limit	67.6	3.93	55	70
Right shoulder width	10.0	1.13	6	14
Left shoulder width	6.82	2.05	4	12
Percent HMA	0.319	0.426	0	1
Mean speed	56.9	2.12	55.9	70.4
Percent of time that the speed was faster than the limit	86.3	16.5	0.81	99.9
Percent of time that the speed was slower than 10 below the limit	0.94	1.87	0	51.7
Percent of time with snowy or icy conditions	0.715	1.21	0	52.9
AADT	37,200	21,544	18,800	111,000
Length (miles)	1.78	1.80	0.15	8.4
Monthly VMT (millions)	79.90	75.64	4.76	466.0
Percent Trucks	30.2	8.00	12.1	40.6
Month is December or January	0.167	0.372	0	1

Table A.2: Crash severity model descriptive statistics (continuous variables)

Variable	Mean	Std. dev.	Min	Max
Mean speed 30 minutes prior to crash	61.7	9.65	6.2	84.5
Standard deviation of speed 30 minutes prior to crash	2.77	3.57	0.00	25.24
Number of Fatalities	0.0137	0.144	0	4
Number of Injuries	0.316	0.718	0	10
Property damage	$17,100	$32,282	$0	$420,000

Table A.3: Crash severity model descriptive statistics (discrete variables)

Variable	
Crash seveirty	O=1,442; B/C=388; K/A=65
Winter weather	"Chemically wet", "icy," or "snowy" = 442; Other = 1,473
Drug/Alcohol-related	At least one driver tested positive or refused drug/alcohol test = 34; Other = 1,861
Weekend	Saturday/Sunday=1,448; Other = 447
Major cause: Swerving	Swerving = 330; Other = 1,565
Major cause: Ran off road	Ran off road = 438, Other = 1,457
Manner of collision: sideswipe	Sideswipe = 717, Other = 1,178
Multiple trucks present	2 or more trucks = 259; 1 truck = 1,638
Multiple non-truck vehicles present	1 non-truck vehicles = 1,070; 0 non-truck vehicles = 825
A truck rear ended a vehicle	Truck rear ended vehicle = 389; Other = 1,506

APPENDIX B. MONTHLY ADJUSTMENT FACTORS

Table B.1: Monthly adjustment factors from average monthly AADT (January–May)

ATR Detector	Year	Jan	Feb	Mar	Apr	May	Jun
103	2011	0.7504	0.7813	0.9208	0.9549	1.0509	1.1225
103	2012	0.7557	0.8036	0.9220	0.9559	1.0677	1.1446
103	2013	0.7619	0.7775	0.9218	0.9369	1.0515	1.1530
103	2014	0.7293	0.7559	0.9143	0.9694	1.0645	1.1481
110	2011	0.7204	0.7460	0.8764	0.8936	1.0744	1.1919
110	2012	0.7458	0.7660	0.9068	0.9303	1.0998	1.2075
110	2013	0.7752	0.7821	0.9327	0.9343	1.0546	1.1819
110	2014	0.7418	0.7697	0.9114	0.9631	1.0688	1.1862
111	2011	0.5555	0.5752	0.6757	0.6890	1.0591	1.2010
111	2012	0.8932	0.9558	0.8426	0.8687	1.0591	1.3309
111	2013	0.7823	0.7944	0.9342	0.9498	1.0640	1.1386
111	2014	0.7416	0.7703	0.9178	0.9672	1.0542	1.1335
115	2011	0.7257	0.7469	0.9048	0.9381	1.0396	1.1495
115	2012	0.7471	0.7619	0.9132	0.9541	1.0786	1.1976
115	2013	0.7441	0.7538	0.9154	0.9179	1.0589	1.1970
115	2014	0.7140	0.7274	0.9118	0.9653	1.0794	1.1911

Monthly adjustment factors from average monthly AADT (January–May continued)

ATR Detector	Year	Jan	Feb	Mar	Apr	May	Jun
116	2011	0.7854	0.8162	0.9186	0.9659	1.0420	1.1159
116	2012	0.8200	0.8333	0.9321	0.9844	1.0650	1.1416
116	2013	0.8058	0.8173	0.9167	0.9481	1.0365	1.1186
116	2014	0.8037	0.8223	0.9225	0.9661	1.0478	1.1254
117	2011	0.8332	0.8788	0.9506	0.9898	1.0349	1.0844
117	2012	0.8582	0.8919	0.9532	0.9971	1.0551	1.1085
117	2013	0.8500	0.8740	0.9501	1.0127	1.0466	1.1217
117	2014	0.8293	0.8627	0.9526	0.9887	1.0437	1.0935
119	2011	0.7415	0.7417	0.8803	0.9461	1.0897	1.1557
119	2012	0.7798	0.8147	0.9272	0.9592	1.0395	1.1436
119	2013	0.8144	0.8168	0.9552	0.9502	1.0617	1.1173
119	2014	0.7414	0.7799	0.9124	0.9863	1.0492	1.1261
120	2011	0.7831	0.8317	0.9277	0.9454	1.0285	1.0366
120	2012	0.7932	0.8463	0.9439	0.9781	1.0584	1.1247
120	2013	0.7905	0.7884	0.9391	0.9349	1.0270	1.1215
120	2014	0.7504	0.7819	0.9147	0.9681	1.0518	1.1262
123	2011	0.7831	0.8134	0.9326	0.9666	1.0441	1.1142
123	2012	0.7978	0.8393	0.9367	0.9686	1.0672	1.1367
123	2013	0.8027	0.8229	0.9395	0.9594	1.0420	1.1461
123	2014	0.7761	0.7995	0.9386	0.9889	1.0749	1.1348

Monthly adjustment factors from average monthly AADT (July–December)

ATR Detector	Year	Jul	Aug	Sep	Oct	Nov	Dec
103	2011	1.1891	1.2099	1.0808	1.0256	1.0231	0.8906
103	2012	1.1675	1.2006	1.0758	1.0242	1.0270	0.8552
103	2013	1.1907	1.2246	1.0476	1.0299	0.9729	0.9317
103	2014	1.1902	1.2115	1.0681	1.0369	0.9974	0.9146
110	2011	1.2227	1.2116	1.0738	1.0286	0.9607	0.8871
110	2012	1.2139	1.2221	1.0930	1.0407	1.0102	0.8755
110	2013	1.2280	1.2048	1.0638	1.0150	0.9348	0.8930
110	2014	1.2263	1.2079	1.0646	1.0302	0.9370	0.8929
111	2011	1.2214	1.2630	1.1126	1.0846	1.0615	0.9532
111	2012	1.3189	1.3823	1.2335	1.1837	1.1965	1.0163
111	2013	1.1704	1.2032	1.0428	1.0313	0.9849	0.9100
111	2014	1.1750	1.2034	1.0616	1.0388	1.0032	0.9333
115	2011	1.2433	1.2461	1.1109	1.0370	0.9829	0.8752
115	2012	1.2177	1.2294	1.0744	1.0097	0.9841	0.8322
115	2013	1.2568	1.2511	1.0752	1.0204	0.9305	0.8791
115	2014	1.2394	1.2097	1.0605	1.0256	0.9688	0.9068
116	2011	1.1532	1.1721	1.0750	1.0452	0.9860	0.9246
116	2012	1.1394	1.1592	1.0413	1.0277	0.9903	0.8656
116	2013	1.1411	1.1943	1.0581	1.0365	0.9881	0.8951
116	2014	1.1446	1.1752	1.0581	1.0365	0.9881	0.8951
117	2011	1.0970	1.1104	1.0396	1.0270	1.0006	0.9537
117	2012	1.0872	1.1076	1.0335	1.0231	0.9985	0.8861
117	2013	1.1046	1.1167	1.0106	1.0154	0.9649	0.9327
117	2014	1.1034	1.1116	1.0386	1.0327	0.9700	0.9734
119	2011	1.1863	1.2057	1.0873	1.0517	1.0002	0.9139
119	2012	1.1430	1.1801	1.0842	1.0402	1.0009	0.8789
119	2013	1.1354	1.1631	1.0490	1.0371	0.9664	0.9333
119	2014	1.1718	1.1768	1.0694	1.0668	0.9842	0.9356
120	2011	1.1639	1.1984	1.0510	1.0307	1.0090	0.9068
120	2012	1.1432	1.1589	1.0248	1.0265	1.0194	0.8826
120	2013	1.1696	1.2103	1.0836	1.0184	0.9852	0.9313
120	2014	1.1788	1.2261	1.0446	1.0540	0.9826	0.9207
123	2011	1.1433	1.1811	1.0605	1.0341	1.0094	0.9176
123	2012	1.1343	1.1705	1.0529	1.0242	1.0066	0.8654
123	2013	1.1570	1.1950	1.0280	1.0238	0.9736	0.9101
123	2014	1.1448	1.1822	1.0471	1.0274	0.9966	0.8977]

BIBLIOGRAPHY

Blincoe, L. J., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R., et al. The economic impact of motor vehicle crashes, 2000. Tech. Rep. DOT HS 809 446, National Highway Traffic Safety Administration, 2002.

Chen, F., and Chen, S. Injury severities of truck drivers in single-and multi-vehicle accidents on rural highways. *Accident Analysis & Prevention*, Vol. 43, no. 5:(2011), pp. 1677–1688.

Dong, C., Burton, M. L., Nambisan, S. S., and Sun, J. Effects of Roadway Geometric Design Features on the Occurrences of Truck-Related Crashes. In *Transportation Research Board 95th Annual Meeting*, 16-1287. 2016.

Duncan, C., Khattak, A., and Council, F. Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions. *Transportation Research Record: Journal of the Transportation Research Board*, , no. 1635:(1998), pp. 63–71.

Federal Highway Administration. Freight Analysis Framework 3. 2011.

Federal Motor Carrier Safety Administration. Hours of service of drivers. Final rule. Tech. Rep. FMCSA 2004 1960, 2011.

Golob, T. F., Recker, W. W., and Leonard, J. D. An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis & Prevention*, Vol. 19, no. 5:(1987), pp. 375–395.

Iowa Department of Transportation. Investigating Officers Crash Reporting Guide. 2014. Accessed Jul. 2, 2015, URL http://www.iowadot.gov/mvd/driverslicense/InvestigatingOfficersCrashReportingGuide.pdf.

Islam, M., and Hernandez, S. Large Truck–Involved Crashes: Exploratory Injury Severity Analysis. *Journal of Transportation Engineering*, Vol. 139, no. 6:(2013), pp. 596–604.

Islam, S., Jones, S. L., and Dye, D. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. *Accident Analysis & Prevention*, Vol. 67:(2014), pp. 148–158.

Kraft, W. H., Homburger, W. S., and Pine, J. L. (eds.). *Traffic Engineering Handbook*. 6th edn. Institure of Transportation Engineers, 2009.

Lord, D., and Mannering, F. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, Vol. 44, no. 5:(2010), pp. 291–305.

Lord, D., Washington, S., and Ivan, J. N. Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, Vol. 39, no. 1:(2007), pp. 53–57.

Martin, J.-L. Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention*, Vol. 34, no. 5:(2002), pp. 619–629.

Mason, J. M., and Smith, B. L. *Accommodation of trucks on the highway: safety in design*. American Society of Civil Engineers, 1988.

Pahukula, J., Hernandez, S., and Unnikrishnan, A. A time of day analysis of crashes involving large trucks in urban areas. *Accident Analysis & Prevention*, Vol. 75:(2015), pp. 155–163.

Savolainen, P. T., Mannering, F. L., Lord, D., and Quddus, M. A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, Vol. 43, no. 5:(2011), pp. 1666–1676.

Stein, H. S., and Jones, I. S. Crash involvement of large trucks by configuration: a case-control study. *American journal of public health*, Vol. 78, no. 5:(1988), pp. 491–498.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. *Statistical and econometric methods for transportation data analysis*. CRC press, 2010.

Wood, G. Generalised linear accident models and goodness of fit testing. *Accident Analysis & Prevention*, Vol. 34, no. 4:(2002), pp. 417–427.

Ye, F., and Lord, D. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: Multinomial logit, ordered probit, and mixed logit. *Transportation Research Record: Journal of the Transportation Research Board*, , no. 2241:(2011), pp. 51–58.

Zhu, X., and Srinivasan, S. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis & Prevention*, Vol. 43, no. 1:(2011), pp. 49–57.